Policy Research Working Paper 9687

Information, Loss Framing, and Spillovers in Pay-for-Performance Contracts

Sebastian Bauhoff
Eeshani Kandpal

WORLD BANK GROUP
Development Economics
Development Research Group
June 2021

Policy Research Working Paper 9687

Abstract

Do incentives matter beyond the information conveyed by pay-for-performance contracts? Does loss framing matter? And do incomplete contracts generate spillovers on unincentivized tasks? This study reports on a framed field experiment with 1,363 maternity care workers in 691 primary health facilities in Nigeria to answer these questions. Participants were randomized into three study arms—(1) information with a flat participation fee, (2) performance-based rewards, and (3) performance-based penalties. In each arm, participants had to identify correct clinical actions based on the records of hypothetical patients receiving maternity care. Five of fifteen possible actions were incentivized but performance was measured on all fifteen. Compared to information alone, both rewards and penalties increase time on task by 11 percent, correct overall performance by 6 to 8 percent, and directly incentivized performance by 20 percent. Incentives also generate positive spillovers of 14 percent on unincentivized tasks. Loss framing does not affect performance. Results suggest that improving health worker effort by 8 percent would have an impact on neonatal mortality at par with the short run effect of adding a physician to a health facility. Finally, findings show that a small incentive captures most of the impact, implying that incentives work by making information more effective and that pay-for-performance contracts can be made significantly more cost-effective.

This paper is a product of the Development Research Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at ekandpal@worldbank.org.

Information, Loss Framing, and Spillovers in Pay-for-Performance Contracts*

Sebastian Bauhoff† Eeshani Kandpal‡

JEL Codes: C93; I11; I15; J41; J45; M52; O15
Keywords: pay-for-performance; effort; crowd out; health workers; field experiment

*We are grateful to counterparts at the Nigerian Federal Ministry of Health and National Bureau of Statistics, Samuel Adebayo, Emmanuel Meribole, and Florence Oke as well as World Bank colleagues Opeyemi Fadeyibi, Benjamin Loevinsohn, Ayodeji Oluwole Odutolu, and Elina Pradhan for their advice and collaboration. We thank Harold Alderman, Jishnu Das, Jonathan de Quidt, Damien de Walque, Quy-Toan Do, Karen Eggleston, David Evans, Madhulika Khanna, Jason Kerwin, Yusuke Kuwayama, Kenneth Leonard, Berk Özler, Owen Ozier, Antonn Park, Bob Rijkers, Ellen Van de Poel, Tom Van Ourti, Adam Wagstaff, and participants of the ASHEcon, MeasureDev, and iHEA conferences for their comments. We appreciate Kevin McGee and Lena Nguyen's support of experiment design and implementation. Mark Conlon provided excellent research assistance. Financial support was provided by the World Bank's Health Results Innovation Trust Fund. The experiment was approved by the Health Media Lab Institutional Review Board (IRB #00001211) and registered with the AER RCT Registry as study AEARCTR-0002482. The findings in this paper are the opinions of the authors, and do not represent the opinions of the World Bank, its Executive Directors, or the governments they represent. All errors and omissions are our own.

†Department of Global Health and Population, Harvard T.H. Chan School of Public Health and Center for Global Development. Email: sbauhoff@hsph.harvard.edu

‡Development Research Group, World Bank. Email: ekandpal@worldbank.org

1 Introduction

Paying for performance (PFP) is a common contracting approach in settings with principal-agent problems (Prendergast, 1999; Lazear, 2000; Duflo et al., 2012; Rothstein, 2015; DellaVigna and Pope, 2018). Such contracts typically provide agents with a checklist of selected outputs and an incentive assigned to each output, and therefore engage agents through two economic channels. First, the checklist explicitly communicates what outputs the principal values and prioritizes. Agents may respond to this information alone, for example, because it reduces their uncertainty about how to allocate their effort or because they are intrinsically motivated (Arrow, 1963). Second, agents may respond to the financial incentives for the specified outputs. They may also respond differently depending on whether the incentives are framed as rewards or penalties (Kahneman and Tversky, 1984; Kahneman et al., 1991; Hossain and List, 2012).

In this paper, we examine the direct and indirect impacts of adding incentives to information on health workers' adherence to clinical guidelines. In an incentivized framed field experiment embedded in a survey of 691 primary care clinics in Nigeria, we randomized 1,363 maternity care workers to three study arms: information, rewards, and penalties. We asked workers to review records, so-called partographs, of five fictitious patients receiving labor and delivery care and to identify clinically necessary—or indicated—actions for each patient. All participants received a participation fee and a checklist of seven common clinical actions related to maternity care of which five were associated with varying prices. Participants in the rewards arm could receive the associated price for correctly identifying the need for the incentivized actions. The penalty contract is isomorphic to the reward contract: we deducted the same amounts for participants who did not recommend the same five actions when they would have been appropriate. We measure performance as the share of indicated actions participants recommend, using the universe of possible clinical actions, which includes actions that were not listed but may be indicated. We assess impacts on overall performance as well as separately for three types of actions: (1) actions that are listed and incentivized in the rewards and penalties arms, (2) those listed but unincentivized, and (3) those neither listed nor incentivized.

Our analysis leverages several key features of the experiment design. First, we randomly listed

or incentivized a subset of all potentially indicated actions but measure performance on the universe of actions. This allows us to examine spillovers on actions that are unlisted or unincentivized. Second, we varied the amounts of incentives across actions, allowing us to estimate price elasticities of effort. Third, we designed the performance-based payout schedule such that the prices and the maximum and minimum payouts are constant across incentive arms. Participants who take the same actions are paid the same in the rewards and penalty arms. Because the two contracts are isomorphic except for the framing, we can isolate loss aversion. In addition, the expected total payout is similar across the arms and participants in all arms receive a participation fee, allowing us to account for an endowment effect.

We find that financial incentives meaningfully improve performance above and beyond information: at similar levels of payouts across the three arms, overall performance is 53 percent in the information arm, 57 percent in the rewards arm, and 56 percent in the penalties arm.[1] Moreover, we find that the incentives crowd-in effort: compared to the information arm, participants in the incentive arms are 3–4 percentage points (pp) more likely to correctly identify unlisted actions. The direct and indirect effects are similar in the rewards and penalties arms, consistent with loss-neutral agents. Finally, our estimate of the price response suggests that a low positive price captures most of the impact of incentives on effort.

Our paper makes several contributions. First, it adds to our understanding of the economic channels through which PFP operates. While a large literature shows that PFP can improve worker performance (Prendergast, 1999), little empirical evidence isolates the relative importance of the information and incentive components of these contracts. This distinction is particularly relevant in health care where workers may be intrinsically motivated (Arrow, 1963; McGuire, 2000; Kolstad, 2013) and perform better in response to measurements that communicate the importance or value of effort, even without incentives (Ashraf et al., 2014; Leonard and Masatu, 2017; Brock et al., 2018; Gauri et al., 2018). Research on PFP in health care in LMICs suggests that conditioning payments can increase performance relative to unconditional payments (Basinga et al., 2011; Hossain and

[1] While a design element of the experiment, described in Section 5, precluded the inclusion of a pure control, we use performance on unlisted actions to benchmark effort in a no-intervention case. In the information arm, performance on unlisted actions (28 percent) is similar to performance on listed actions (26 percent), suggesting that providing information alone may not increase performance.

List, 2012; Diaconu et al., 2020), but it is unclear whether the effect can be attributed to the incentive channel alone. Our experiment resolves the practical challenge that incentives always also convey information because PFP contracts must state what is incentivized.

Second, we provide a conceptual framework and empirical evidence on spillovers in incomplete PFP contracts. A long-standing concern with PFP schemes is that incomplete contracts, i.e. those that only incentivize a subset of actions, could lead agents to "multitask" by diverting effort toward actions associated with the highest net gain (i.e., incentive net of cost), possibly at the expense of actions that have a lower payoff or are unmeasured (Holmstrom and Milgrom, 1991; Prendergast, 1999; Mullen et al., 2010; Miller and Babiarz, 2013; Finan et al., 2015; Bulte et al., 2021). Alternatively, performance-based contracts may crowd in effort, even on actions that are unincentivized or have a relatively low gain, for example, because of complementarities in production (Mullen et al., 2010; Sherry, 2016). The limited available literature does not find evidence of spillovers of PFP in health care, even for large schemes such as the United Kingdom's Quality and Outcomes Framework PFP scheme (Campbell et al., 2007) or national PFP schemes in LMICs (Sherry et al., 2017; Celhay et al., 2019). We show theoretically that the magnitude and direction of spillovers will depend on the degree of complementarity in production. Further, we address a practical challenge in detecting spillovers, which is that workers can adjust effort on many—possibly unobserved—margins by using well-defined and self-contained experimental tasks for which we can assess performance on the universe of relevant actions. Our findings suggest that providing incentives for some actions also increases effort on unlisted actions by 4 pp, or 14 percent, which is indicative of complementarities in production.

Third, our paper also contributes to research contrasting the effects of loss or gain framing in contracts. This distinction is scientifically and practically important, as penalties can be politically and logistically challenging to implement because they involve withholding or recovering payments. Some experimental evidence—including from LMICs—suggests that identical incentives can be more effective if cast as penalties for poor performance rather than as rewards for good performance because of loss aversion (Kahneman and Tversky, 1984; Kahneman et al., 1991; Fryer et al., 2012; Hossain and List, 2012; Imas et al., 2017; Bulte et al., 2020, 2021). In contrast, other research suggests that framing does not matter very much in general settings (de Quidt et al., 2017; de Quidt,

2018; DellaVigna and Pope, 2018) and that loss framing in the form of penalties can lead to negative spillovers through gaming behaviors (Pierce et al., 2020). When we contrast the two frames for the same experimental task and isomorphic contract, we find that rewards and penalties have similar direct and spillover effects. This is indicative of a loss-neutral agent.

Finally, our findings can inform the design of PFP contracts in health care. PFP has long been used in health systems in high income contexts (Campbell et al., 2007; Mendelson et al., 2017) and is increasingly deployed in LMICs, where the poor quality of health care services stems partly from low effort by health workers (Das and Gertler, 2007; Leonard et al., 2007; Das et al., 2008; Leonard and Masatu, 2010). Penalties are not common but nonetheless used, for example, in the United States' Medicare's Nonpayment Program, which withholds reimbursements for costs related to hospital-acquired conditions with the goal of reducing the incidence of these conditions (Gupta, 2021). In our context, our findings provide a rationale for adopting at least small incentives in lieu of information-only interventions, such as job aids, the dissemination of guidelines, or training programs (Rowe et al., 2005). They also suggest that, in practice, there may be little cost to implementing the simpler and more palatable rewards frame in real-world PFP contracts.

Our experimental task is hypothetical—participants recommend actions without performing them—and there are no serious consequences of poor performance, such as harm to patients. However, the experiment design and context are realistic (Prendergast, 1999; Harrison and List, 2004). In particular, we used meaningful incentives, the task mimics what participants do routinely in their jobs, and the study sites are their primary workplaces. We also find that participants' performance and response patterns align with behavior observed in real-life primary health care provision in LMICs. As in our experiment, other research has found that health workers perform only about half of the clinically appropriate actions and that they might perform actions that are unnecessary and potentially harmful to patients (Das et al., 2008, 2016; Lopez et al., forthcoming). Our performance measure is significantly and positively correlated with separate assessments of participants' knowledge and clinical practice that were conducted alongside our experiment. Moreover, our study participants appear to take the task seriously and the response pattern is not consistent with a mechanical response to the checklist or incentives, for instance, as participants in the information arm exert some effort even if they do not stand to gain financially from it. In addition,

the difference in performance across arms is largest for the middle of the performance distribution rather than the bottom of the distribution, suggesting that the incentives did not only affect participants who would have not paid any attention otherwise. Similarly, participants frequently identified actions that are indicated but not listed or incentivized. Taken together, these findings suggest that the participants paid attention and exerted effort rather than simply identifying the listed or incentivized actions.

The rest of this paper proceeds as follows. Sections 2, 3, and 4 present a conceptual framework, the experimental design, and the data used, respectively. Section 5 discusses methods, including our definition of performance outcomes. Section 6 presents the results and discusses the validity of our measure, optimal contracts, and cost-effectiveness. Section 7 concludes.

2 Conceptual framework

In this section, we sketch out a framework to guide our empirical analysis of how the information and incentive channels of PFP operate and interact. We generalize DellaVigna and Pope's (2018) model by (1) considering an agent who is optimizing effort allocation between multiple clinical actions, (2) making the returns to motivation or recognition a function of information, and (3) considering cross-price effects, that is, the impact of one action's incentive on the effort allocated to another action. Spillovers in the rewards and penalties arms can be negative or positive: on the one hand, multitasking may increase effort on incentivized actions and reduce effort on unincentivized ones, while on the other hand, some actions may share common inputs or processes so that effort on one action may increase output on others. For exposition, we do not consider more complex issues, such as interactions among multiple incentivized actions (Mullen et al., 2010; Sherry, 2016).

Beginning with the rewards arm, consider the risk-neutral agent's optimization problem when facing a flat participation fee, Π_r, and two actions that each are associated with a non-pecuniary "reward," s, which is a function of information about that action, i, and a price, r, that is paid to

the agent if she performs the action:

$$\max_{e_1 \geq 0, e_2 \geq 0} \Pi_r + [s(i_1) + r_1]e_1 + [s(i_2) + r_2]e_2 - c(e_1, e_2). \tag{1}$$

We assume a convex cost of effort function, $c(e)$; that is, $c'(e) > 0$ and $c''(e) > 0$ for all $e > 0$. Optimal effort e^* is then increasing in both the non-pecuniary and per-unit pecuniary rewards. First-order conditions can be written as

$$s(i_1) + r_1 - \frac{\partial c(e_1^*, e_2^*)}{(\partial e_1)} = 0, \tag{2}$$

$$s(i_2) + r_2 - \frac{\partial c(e_1^*, e_2^*)}{(\partial e_2)} = 0. \tag{3}$$

Second-order conditions can be written as

$$\frac{\partial^2 c(e_1^*, e_2^*)}{\partial e_1^2} \geq 0, \tag{4}$$

$$\frac{\partial^2 c(e_1^*, e_2^*)}{\partial e_2^2} \geq 0, \tag{5}$$

$$\left[\frac{\partial^2 c(e_1^*, e_2^*)}{\partial e_1 \partial e_2}\right]^2 - \frac{\partial^2 c(e_1^*, e_2^*)}{\partial e_1^2} \frac{\partial^2 c(e_1^*, e_2^*)}{\partial e_2^2} \leq 0. \tag{6}$$

Taking total derivatives of equations (2) and (3) with respect to r_1,

$$-\frac{\partial^2 c}{\partial e_1^2} \frac{\partial e_1^*}{\partial r_1} - \frac{\partial^2 c}{\partial e_1 \partial e_2} \frac{\partial e_2^*}{\partial r_1} + 1 = 0, \tag{7}$$

$$-\frac{\partial^2 c}{\partial e_1 \partial e_2} \frac{\partial e_1^*}{\partial r_1} - \frac{\partial^2 c}{\partial e_2^2} \frac{\partial e_2^*}{\partial r_1} = 0. \tag{8}$$

Equation (8) can be rewritten as

$$\frac{\partial e_2^*}{\partial r_1} = -\frac{\frac{\partial^2 c}{\partial e_1 \partial e_2} \frac{\partial e_1^*}{\partial r_1}}{\frac{\partial^2 c}{\partial e_2^2}}. \tag{9}$$

Plugging equation (9) into equation (7), we get the following expression for the response of optimal effort on an action to its own price:

$$\frac{\partial e_1^*}{\partial r_1} = -\frac{\frac{\partial^2 c}{\partial e_2^2}}{\left(\frac{\partial^2 c}{\partial e_1 \partial e_2}\right)^2 - \frac{\partial^2 c}{\partial e_1^2}\frac{\partial^2 c}{\partial e_2^2}}. \tag{10}$$

We know from the second-order conditions that the denominator on the right-hand side of equation (10) is negative. From our assumption of a convex cost function, $\frac{\partial^2 c}{\partial e_2^2} > 0$. Thus, we have $\frac{\partial e_1^*}{\partial r_1} > 0$, meaning that, holding information constant, providers increase effort allocated to an action in the price of that action. Plugging this into equation (9) gives us the following expression for $\frac{\partial e_2^*}{\partial r_1}$:

$$\frac{\partial e_2^*}{\partial r_1} = \frac{\frac{\partial^2 c}{\partial e_1 \partial e_2}}{\frac{\partial^2 c}{\partial e_1 \partial e_2}^2 - \frac{\partial^2 c}{\partial e_1^2}\frac{\partial^2 c}{\partial e_2^2}}. \tag{11}$$

We know from equation (10) that the denominator is negative, so if actions are complements, the sign of $\frac{\partial^2 c}{\partial e_1 \partial e_2}$ is negative and we have $\frac{\partial e_2^*}{\partial r_1} > 0$. On the other hand, if actions are substitutes, $\frac{\partial^2 c}{\partial e_1 \partial e_2}$ is positive and we have $\frac{\partial e_2^*}{\partial r_1} < 0$. Intuitively, if actions are completely unrelated, effort on an action is independent of the price of other actions.

In the penalty arm, the provider's optimization problem with a flat participation fee, Π_p, two actions, and penalties, p, is

$$\max_{e_1 \geq 0, e_2 \geq 0} \Pi_p + s(i_1)e_1 - \lambda(\bar{e_1} - e_1)p_1 + s(i_2)e_2 - \lambda(\bar{e_2} - e_2)p_2 - c(e_1, e_2), \tag{12}$$

where λ is a parameter of loss aversion such that a loss-averse individual has $\lambda > 1$, while a loss-neutral individual has $\lambda = 1$. The first-order and second-order conditions are analogous to those for a positive price for effort.

Solving for own- and cross-price elasticities of effort yields the following expressions:

$$\frac{\partial e_1^*}{\partial p_1} = -\lambda\frac{\frac{\partial^2 c}{\partial e_2^2}}{\left(\frac{\partial^2 c}{\partial e_1 \partial e_2}\right)^2 - \frac{\partial^2 c}{\partial e_1^2}\frac{\partial^2 c}{\partial e_2^2}}, \tag{13}$$

and

$$\frac{\partial e_2^*}{\partial p_1} = \lambda \frac{\frac{\partial^2 c}{\partial e_1 \partial e_2}}{\left(\frac{\partial^2 c}{(\partial e_1 \partial e_2}\right)^2 - \frac{\partial^2 c}{\partial e_1^2}\frac{\partial^2 c}{\partial e_2^2}}. \tag{14}$$

For a loss-neutral person, equations (13) and (14) for penalties are identical to equations (10) and (11) for rewards. Thus, in the absence of loss aversion, workers choose the same optimal levels of effort in response to a reward or an equivalent penalty. In contrast, loss aversion would imply that a given increase in the penalty on action 1 leads to an increase in effort on action 2 when the actions are complements, and a decrease in action 2 when they are substitutes.

In sum, our model predicts that, holding information constant, incentives should increase effort on the incentivized actions and the degree of complementarity between actions determines the sign of any spillovers between actions. In particular, incentives on one action will raise effort on complementary actions (positive spillover) but decrease efforts on actions that are substitutes (negative spillover). Finally, loss-averse agents are more responsive to a penalty than an equivalent reward. The direction and degree of complementarity between actions and the degree of loss aversion are empirical questions that we examine below.

3 Experiment design

We used a framed incentivized field experiment to empirically examine the direct and spillover effects of adding incentives to information. In brief, during a survey of maternity care workers in Nigeria, we asked 1,363 respondents in 691 health facilities to participate in a study relating to adherence to protocol in the context of labor and delivery care. Our experimental task is aligned with the study participants' day-to-day work: deciding on clinical actions that are appropriate for women receiving labor and delivery care.

3.1 Experimental task

We first asked participants to review records for five fictitious patients in different stages of labor and then to identify all the clinical actions that are appropriate for each patient. The clinical record—called a partograph—is a standard tool that is "strongly recommended" by the World Health Organization and used by Nigeria's Federal Ministry of Health to train health workers providing maternal health services (WHO, 2014; White Ribbon Alliance, 2015). The correct use of partographs is also often part of PFP programs for health care in LMICs, although not in the Nigerian trial that our experiment was embedded in (Fritsche et al., 2014). We designed five partograph cases based on examples from medical training materials and hired two medical professionals to independently identify actions that would be clinically indicated or unnecessary based on standard clinical guidelines in the management of labor and delivery. Unnecessary actions at best serve no medical purpose and may even be potentially harmful to the patient.

We fielded two types of partograph cases that represent typical scenarios in our settings. In the two "simple" cases, we asked participants to assess whether a single action suggested by an unnamed colleague is correct or incorrect. In the three "complex" cases, the respondent is described as being in charge of the patient and is asked to name all actions that she deems clinically indicated for that patient. Table A.1 presents all possible actions for the three complex tasks and notes whether a given action is indicated or unnecessary. Which actions are indicated varies across cases but is invariant across participants. We scored as "correct" actions that participants correctly identified as indicated as well as those that participants did not identify as indicated and are, in fact, clinically unnecessary.

Table 1 presents the prices and proportion of correct responses for each action, disaggregated by whether the action is listed with or without incentive or is unlisted.[2] There is considerable variation in performance across actions as well as an overlap in the range of performance for listed and unlisted actions. We observe relatively high levels of performance for many listed and unlisted actions. For example, health workers correctly identify the need for referral to a higher-level facility (an incentivized action) 75 percent of the time and correctly recommend administering magnesium

[2] Table A.2 disaggregates performance on each of these actions for the three complex cases.

sulfate (an unlisted action) 96 percent of the time. In contrast, participants often missed other actions that are always appropriate: monitoring contractions and the amniotic fluid, tracking the fetal heart rate and the mother's vital signs, and recording the fluids and drugs administered. Performance is also generally higher for clinically unnecessary or incorrect actions. For example, participants almost never recommended administering magnesium sulfate or augmenting labor, which are unnecessary in all cases.

3.2 Study arms

The experiment was embedded in a survey of primary care clinics fielded as part of an impact evaluation of health facility financing modalities (see Section 4.1 below). Toward the end of the survey, we randomized respondents in each clinic into one of three trial arms (information, rewards, penalties), stratified by clinic. All participants received a printed list of seven randomly selected common clinical actions that are shown in Table 1, for example, to monitor the fetal heart rate or prepare for imminent delivery. The list shown to participants is purposively incomplete: there are eight additional actions not on the list but may be appropriate. In the rewards arm, participants were offered payments if they recommended five of the seven actions when they were indicated for the patient.[3] In the penalties arm we deducted the same payment amount for each action that is indicated but was not identified by the participant. The reward and penalty contracts are isomorphic: the same actions lead to the participant being paid the same in either arm. We randomized the payment amounts across the incentivized actions, and the listed and incentivized actions include some that are unnecessary or even harmful in some cases.

All participants received a flat payment that varied by arm such that the maximum, minimum, and, in expectation, average payouts are all the same. The flat payment in the information arm was 1,750 Nigerian Naira or about USD 5.80. The participation fee allows us to account for an endowment effect, which may be important because most public sector contracts have fixed remuneration scales and PFP interventions at scale aim to be budget neutral (Fritsche et al., 2014).

[3]While we randomly selected actions to be listed, due to the small overall number of actions, the three groups of actions (listed, paid, unlisted) could still be systematically different. We address this issue below.

The rewards group received a base fee of 1,000 Naira (USD 3.30) and could gain up to 2,500 Naira (USD 8.30), while the penalties group received a base pay of 2,500 Naira and could lose up to 1,500 Naira (USD 5) for a minimum payout of 1,000 Naira.[4] The incentives for individual actions in the rewards and penalties arm varied between 50 and 300 Naira (USD 0.17 to 1). We compensated participants in the form of cellphone airtime after they had completed all five cases. Appendix B presents the instructions provided to participants, payout schedule, screenshots of representative interview templates, and all partographs.

3.3 Key design features

The experimental design allows us to achieve four objectives. First, we can examine (across arms) the effect of adding incentives to information, as participants in the information arm only received the list of seven actions, while those in in the rewards and penalties arms were also offered incentives for five of these actions. This allows us to disentangle the effects of information and incentives in PFP contracts. Second, we can examine spillovers between types of action (listed, incentivized, unlisted). Third, we can study the relative effect of gain and loss framing by comparing the rewards and penalty arms within each action type. Finally, we can use the variation in incentive amounts across actions to examine price responses.

Our experimental setup did not permit us to assess the effect of information relative to a pure control in which participants are not even provided with the list of possible actions. This is because the partograph could convey information. For instance, it provides space to track the mother's vital signs and is designed to guide clinical decisions via alert and action lines that trigger emergency referrals or procedures, such as initiating a cesarean section. As discussed below, in practice the two actions that are not listed but are mentioned on the partograph—measuring descent and mother's vital signs—are rarely named correctly, suggesting that the information conveyed by the partograph is not particularly salient. Below we describe how we use performance on unlisted actions to benchmark effort in a pure control condition.

[4] One possible concern is that our loss framing failed to change participants' reference points by very much. While we did not prepay the participation fee in the penalty arm, the instructions explicitly stated that participants stood to lose part of their participation fee.

Finally, Appendix C notes two minor adjustments to the secondary analysis we had pre-registered as AEARCTR-0002482. First, we cannot analyze balance by facility catchment size and health worker education level because the data on catchment size are largely missing (not reported at the facility level) and because there is little variation in health worker education level. We explore tenure and experience in lieu of education and find the arms to be balanced. Second, we reassigned midwives to the medical professional rank. We had intended to include them as lower-level professionals, but found that their training, experience, and salary are comparable to that of a nurse. The results are robust to their inclusion in lower-level ranks instead.

4 Setting and data

In this section, we describe the study setting and data collection as well as the outcome measures, key covariates, effect moderators, and potential confounders.

4.1 PFP trial and impact evaluation

We embedded the experiment in the endline survey of a concurrent cluster-randomized trial of different health facility financing modalities in Nigeria (Kandpal et al., 2019). This trial randomized all 52 districts in three states to two arms. A total of 1,389 primary and secondary care facilities were either assigned (1) to PFP with quarterly bonuses based on the quantity and quality of primary health services they provided or (2) to direct facility financing (DFF) that disbursed half of the average PFP bonus without conditioning the payment on performance. In both arms, district supervisors administered a checklist to assess quality of care on a quarterly basis; an independent agency verified performance in the PFP facilities. A "business as usual" control group was established by selecting three states in the same geopolitical zone that border the intervention states and resemble them along observed demographic dimensions that are associated with primary health care outcomes. Appendix Figure A.1 shows the intervention states (Adamawa, Nasarawa, Ondo) and the control states (Taraba, Benue, Ogun) as well as the locations of the health facilities in this study.

For the impact evaluation of the concurrent trial, one primary or secondary health facility was randomly chosen per ward in each of the districts, for a sample of 786 facilities out of the 1,389 facilities participating in the trial. At each facility, two health workers who routinely provide antenatal or under-five curative care from the roster of health workers present on the day of the survey were randomly sampled for an in-depth interview. We conducted our experiment in all 691 primary health facilities in the survey sample of the impact evaluation. All health workers in a selected facility were eligible to participate in the experiment. The experiment was conducted simultaneously with participating health workers in the same facilities being interviewed in different rooms to prevent information sharing.

The survey instrument captured a range of health worker characteristics, including their education, when they have been last trained in labor and delivery, and their professional grade. In addition, the survey included two standard assessments of worker knowledge and skill (see for instance, Das and Gertler, 2007) in the preceding non-experimental section. First, it included a protocol-based vignette based on the standard WHO antenatal care protocol (Villar et al., 2001) in which participants were read a narrative about a pregnant woman seeking antenatal care and were asked to list everything that they would do during that visit.

Second, the survey provides a measure of real effort—i.e. one with costs to the health worker—through direct observations of actual patient-provider interactions in the context of antenatal care. As described in Kandpal et al. (2019), direct observations were performed at a randomly-chosen third of the sampled primary health centers. In each health center, one randomly selected health worker was observed while she was providing antenatal care to two patients. We thus have direct observation data for 339 of the 1,363 health workers in our experiment. The observation was conducted using a structured, quantitative checklist and the data were collected by enumerators trained in the direct observation of antenatal care provision. Enumerators recorded whether the health worker performed actions listed in the standard WHO protocol, and whether she performed five standard screening tasks to assess the woman's risk for developing serious pregnancy complications. The latter are closely related to our experimental task of identifying emerging complications and guiding interventions. Moreover, like in our task, the screening can be performed in any setting whereas the physical actions may require supplies or equipment, such as a stethoscope or lab kits for

blood tests. We therefore use only the screening tasks to construct three measures of performance: whether the worker performed any of the screening actions (60 percent of participants), whether she performed all five (22 percent), and the total number of actions she performed (two of five, on average).

The PFP trial started in July 2014, when the two financing interventions were rolled out. The evaluation endline survey that contained our experiment was conducted between August and October 2017. The impact evaluation found that districts with PFP or DFF performed better than those in the control group and the impacts of PFP and DFF were comparable, with few exceptions (Kandpal et al., 2019). Both PFP and DFF had a practically and statistically significant impact on the quantity of key maternal and child health services. For example, both significantly increased fully immunized child coverage and modern contraceptive prevalence. However, directly observed clinical quality of care, which may be most directly related to provider effort, showed limited gains, especially in the DFF arm.

For context, 75 percent of all health workers in our sample reported working seven days a week, for an average of six hours a day. The median monthly gross salary is 43,000 Nigerian Naira or about USD 113. Wages are stagnant and often paid with a delay: only a third of health workers reported receiving a salary increase in the last two years, and a quarter said they had not received their full pay for the previous month. Indeed, 63 percent reported not having received their entire salary for the past year.

4.2 Outcomes, moderators, and possible confounders

Our measure of overall performance is the proportion of clinical actions that participants correctly identified as medically appropriate for each partograph case. Since actions can be clinically indicated or not, we score as correct those actions that are indicated and named by respondents as well as actions that are unnecessary and not named.[5] In other words, a fully correct set of actions would be one in which the participant identifies all indicated actions and none that are unnecessary. We

[5]In four instances, actions can be ambiguous based on the partograph Table A.1. Because performing ambiguous actions can be unnecessary or harmful to patients, we consider these actions to be not indicated.

then calculate the proportion of correct actions as a share of all possible actions. However, note that we have simple tasks and complex partograph cases. In the simple cases, respondents must merely identify whether a single action framed as a suggestion from a peer is correct based on the information provided in the partograph. In contrast, for complex cases, respondents must list all the appropriate actions. To account for this difference in the number of relevant actions across cases, we measure overall performance in two ways: weighting the cases equally or weighting the responses equally. Specifically, in the "across cases" measure, we calculate the proportion separately for each of the five cases and then average across cases. In the "across responses" measure, we calculate the proportion correct for actions in all cases.

We consider two possible confounders. First, we assess the robustness of our treatment impacts to the inclusion of participation in the arms assigned in the concurrent cluster-randomized trial: control, PFP, or DFF. Participants in the PFP arm may have been comparatively more attuned in responding to incentives (Leaver et al., forthcoming). Further, Kandpal et al. (2019) show that awareness of the PFP was a significant mediator of the its effectiveness.[6] We therefore also examine the effect of a binary measure of respondents' self-reported awareness that their clinic is participating in the ongoing program on their performance in the experimental task. Second, to address the concern that we measure knowledge or skill rather than effort, we examine whether the responses to our task are explained by the health worker's knowledge of the maternal care protocol or their skill as measured by the direct observations. We calculate this level of knowledge as the share of actions or screening tasks the respondent correctly identified in the above-mentioned antenatal care vignette or in the direct observation, respectively, and use a binary indicator of whether the participant scored above the sample median.

Table 2 shows summary statistics for the outcomes and key covariates across the study arms. There are about 450 participants in each arm. Overall performance is low: on average in the information arm, only about half of participants' responses are correct, that is, either indicated when clinically indicated and not identified when unnecessary. The average payout in the rewards and penalties arms is comparable and slightly higher than in the information arm. Although

[6]However, that even in the health facilities assigned to either the PFP or DFF trial arms, a majority of health workers had either not heard of the trial or did not understand its structure.

treatment assignment was randomized, there is some imbalance. In particular, participants in the information arm are 5.4 pp more likely to be male than in the rewards arm and 6.8 pp more likely to be male than in the penalty arm. In addition, participants in the rewards arm are 3.6 pp less likely to be doctors, nurses, or midwives compared to in the information arm. We assess the potential impacts of the observed imbalance in sex and grade below.

4.3 Validity of the experimental task and performance measure

Although the experiment revolves around fictitious patients and does not impose actual effort costs on health workers—participants only identify appropriate actions but do not actually implement them on patients—there are several design features that may help make this setup realistic (Harrison and List, 2004). In particular, the incentives are real, the participants are actual health workers whose daily work—providing labor and delivery care—aligns with our experimental task, and the study was conducted in their primary workplace. We also find that overall performance on our task and measure is comparable to non-experimental assessments of knowledge and actual performance by the same health workers. Specifically, participants in the information arm have an overage score of 53 percent on our task, which is similar to the average scores on the knowledge vignette (about 53 percent) and the screening tasks of the direct clinical observations (60 percent) for the full sample. This level of quality of care is typical for LMIC settings (Das et al., 2008, for example).

Further, the response patterns indicate that participants took the experimental task seriously and exerted effort. First, in the two simple cases, participants respond yes or no. If they were randomly selecting a response, we would expect to see responses of approximately 50 percent for each of these cases. Instead, we observe 69 and 27 percent correct performance, respectively, which suggests that we are capturing actual variation even in the so-called simple cases. Second, participants in the information arm exerted effort even if they did not stand to gain from it. Third, we find that the gains from incentives come from the middle of the performance distribution which suggests that impact of incentives is not driven by participants who were not paying any attention at all.

Fourth, participants in the incentive arms did not merely minimize effort by naming all paid

actions, although doing so would have only increased their payout (with the exception of unnecessary referrals). Fifth, as highlighted in Table 1, participants identified both unnecessary actions that were paid for (mimicking real-life overuse) as well as actions that were neither listed nor incentivized (for example, measuring vital signs), suggesting they did not simply respond by naming the listed or paid actions. Sixth, we find that participants in the incentive arms spend more time on the interview than those in the information arm (Table A.3), which is consistent with increased effort. In other settings, increased time spent is associated with improved performance (Das et al., 2012; Rivkin and Schiman, 2015; Lavy, 2016; Cattaneo et al., 2017).[7]

Seventh, our estimated price elasticities are relatively low and consistent with estimated wage elasticities in LMICs (Goldberg, 2016). In the context of hypothetical tasks with limited effort costs, if anything, one might expect artificially high price elasticities; instead our elasticity estimates are comparable to "real world" estimates. Finally, we find that our results are robust to controlling for participants' knowledge and costly effort as measured with the antenatal care vignette and direct observations, respectively. This suggests that participants took the experiments seriously and that our task induced real attention and effort.

5 Empirical estimation

We conduct our analysis in five steps. First, we estimate overall performance across all actions for each of the three treatment arms. Then, we examine the effect of incentives by comparing performance on the subset of five incentivized actions across arms. Third, we assess incentive spillovers by comparing performance across arms on listed and unlisted actions. Fourth, we examine the effects of the two potential confounders on overall performance: whether the facility was assigned to PFP, DFF, or control in the larger trial and whether the participant had a higher than median score on the knowledge test of maternity care protocol. Finally, we benchmark the effect of information

[7]Some interviews were not completed within the day that they were started and the survey enumerator returned to the facility when the health worker was next on shift to complete the interview. There may also have been instances in which an enumerator failed to promptly record the interview as complete. In these instances, the interview length cannot be calculated correctly from time stamps because the end date was several days after the start, and the interviews were not "paused" in the interim. We omit these observations from the analysis of time-on-task. This leaves us with a sample of 1,178 observations.

relative to a pure control.

We leverage the randomized assignment and estimate variants of the following OLS model:

$$y_{if} = \alpha + \beta \cdot Incentives_{if} + \gamma_f + \eta_{if}, \tag{15}$$

where y is the performance of participant i in facility f and the vector $Incentives$ captures whether they were randomly assigned to rewards or penalties. Standard errors are clustered at the facility level. We assess robustness with additional regressions that control for the participant-level covariates listed in Table 2, including interactions of the treatment indicators with the two unbalanced covariates, gender and job grade.

Overall performance by arm: We begin by assessing overall performance in the three study arms: information, rewards, and penalties. These estimates capture the effect on all possible clinical actions, that is, those that are unlisted, listed, or paid. This captures the combination of direct effects arising from financial incentives and information and indirect effects arising from spillovers on unincentivized or unlisted actions. Specifically, we calculate the following differences:

1. (Rewards Arm, All Actions – Information Arm, All Actions)
2. (Penalties Arm, All Actions – Information Arm, All Actions)

Effect of adding incentives to information: Next, we study the impact of incentives above and beyond information by contrasting performance on actions that are paid in the incentive arms with performance on the same actions in the information arm. For the two incentive arms, paid actions necessarily contain both incentives and information. We subtract the performance in the information arm (pure information) from the performance in the incentivized arms (incentives and information). This yields the effects of pure incentives (i.e., incentives net of information) on paid actions in the presence of possible spillovers:

3. (Rewards Arm, Paid Actions – Information Arm, Listed Actions)

4. (Penalties Arm, Paid Actions – Information Arm, Listed Actions)

Spillover effect of incentives on listed and unlisted actions: Third, we examine the spillover effects of incentivizing a subset of actions on performance on actions that are merely listed (but not incentivized) or are even unlisted. Specifically, we compare performance in the rewards and penalties arm (where spillovers from incentives could exist) to performance in the information arm, for listed and unlisted actions. This yields our estimate of the incentive spillovers relative to information alone:

5. (Rewards Arm, Unlisted Actions – Information Arm, Unlisted Actions)

6. (Penalties Arm, Unlisted Actions – Information Arm, Unlisted Actions)

7. (Rewards Arm, Listed Actions – Information Arm, Listed Actions)

8. (Penalties Arm, Listed Actions – Information Arm, Listed Actions)

Potential confounders of overall performance: We examine the role of two potential confounders for overall performance by interacting our three study arms with $Confounder_{if}$. First, we interact assignment to an experimental arm with a vector of indicators for the assignment in the larger cluster-randomized trial to PFP, DFF, or control. This allows us to examine whether workers who are exposed to the larger PFP might generally exert more effort (Leaver et al., forthcoming). While such an effect would not invalidate our effect estimates because we randomized within facilities, it could inflate our estimates of the impact of incentives on the relevant subsample and limit external validity. Second, we interact our study arms with a binary indicator of whether the participant scored above the median on the knowledge test. The estimation equation is as follows:

$$y_{if} = \alpha + \beta \cdot Incentives_i + \kappa \cdot Confounder_{if} + \gamma \cdot Incentives_{if} \cdot Confounder_{if} + \eta_{if}. \quad (16)$$

Benchmarking performance against no intervention: Finally, as noted in Section 3, we cannot estimate the effect of information relative to a pure control. In particular, we did not design the experiment to include a pure control arm because in principle the partograph itself could convey some information. In practice, the two items that are not listed on the checklist but are mentioned in the partograph—measuring descent and mother's vital signs—are least likely to be correctly mentioned, as Table 1 shows. We can obtain a rough benchmark for a pure control by comparing, in the information arm, performance on listed actions with performance on unlisted actions. There are two important caveats: first, the actions in these two groups may not be comparable, and therefore performance could be different in the absence of the information we provided. Second, there could be spillovers onto these "pure control" actions, for instance, if participants in the information arm shift effort toward actions on the checklist. If there are negative spillovers from information on unlisted actions, our estimate of the effect of information would be an upper bound, while a positive spillover would lead to a lower-bound estimate. With these issues in mind, we can estimate the effect of information relative to a pure control as:

9. (Information Arm, Listed – Information Arm, Unlisted)

6 Results

In this section, we first discuss findings related to the direct effects of incentives on overall performance and on actions associated with incentives in the rewards and penalties arms. Then, we turn to the estimation of spillover effects on actions that are unlisted or listed but unpaid. Next, we rule out two key confounders: exposure to the larger PFP trial and the participant's clinical knowledge. We conclude with the benchmarking of the effect of information against no intervention.

6.1 Effect of adding incentives to information

We start by examining the effects of adding rewards or penalties to information on the distribution of overall performance and on average performance. We focus the discussion on the "across cases"

performance measure that weights cases equally; however, the findings are similar for the "across responses" measure that weights responses equally. A comparison of the empirical cumulative distributions of performance in Figures 1a and 1b, equally weighing cases and responses, respectively, yields three findings. First, the range of observed performance is comparable across all arms. Second, the two incentive arms perform substantively and statistically better than the information arm (Kolmogorov-Smirnov tests p=0.00) largely from a shift in the middle of the distribution. Third, the distributions of the rewards and penalty arms are not economically or statistically significantly different.

We observe the same pattern in the regression results for average effects presented in Table 3. Performance in the information arm is 53 percent and 4.3 and 3.4 pp higher across cases in the rewards and penalties arms, respectively, and approximately 2 pp higher across responses. This suggests that overall, incentives increase average performance by 6 to 8 percent. These results are robust to the linear and interacted inclusion of covariates, specifically gender and job grade (Tables A.10 and A.11). For both measures, the rewards and penalty arms are statistically indistinguishable in terms of impacts on performance, with a p-value of 0.29 reported in the bottom panel of Table 3. Such a lack of difference between rewards and penalties is consistent with an agent who is loss neutral, as discussed in Section 2.[8]

As not all the cases are structured in the same way, Columns (3)–(7) of Table 3 report the impact estimates for each case separately. By and large, we find that the rewards and penalties arms outperform the information arm for each case and there is no statistically significant difference between performance in the rewards and penalty arms. The different structure of the assessment across the cases also appears to matter, as the two simple cases have higher impacts of rewards, with 5.3 pp (27 percent) and 11 pp (69 percent), compared with magnitudes of 1 to 3 pp (3 to 5 percent) for the complex cases. Even though the task structure is simple—providing a yes or no answer—it is not necessarily the case that arriving at that answer is trivially easy. Indeed, performance in the information arm is at both its highest and lowest for the two simple cases, at

[8] Table A.4 reports impacts on overall performance measured in z-scores instead of proportion correct. Our PAP, discussed in Appendix C, lists both measures of overall performance, and thus we report both. The results are robust to either way of measuring performance. However, given that the two types of tasks—simple and complex—involve different numbers of actions and thus have significantly different underlying distributional variation, we prefer the proportion of correct responses as an outcome measure.

69 percent and 27 percent.

Since our measure of overall performance includes actions that were paid and those that were unpaid, we next examine the effects of incentives on the subset of actions that were paid. We conduct this analysis at the level of individual actions from the complex cases. Table 4 reports the average performance on all three types of actions (unlisted, listed, and paid) by arm; the full regression results are presented in Table A.5. Focusing on paid actions, we find evidence for direct effects of incentives on paid actions. Performance in the two incentive arms is 7.4 (rewards) to 7.9 (penalties) pp higher than for the same actions in the information arm. For unnecessary actions that are paid, performance in the penalty arm is comparable to the information arm but is 2.9 pp higher in the rewards arm. In neither case do we find detectable differences between rewards and penalties. For these same actions, performance in the information arm is 38.1 percent for indicated actions and 79.4 percent for unnecessary actions, suggesting that adding incentives to information increases performance by about 20 and 3.5 percent, respectively. In all arms, performance on the unnecessary actions is higher than for indicated actions.

In Table 4, we also distinguish between clinically indicated and unnecessary actions. Notably, we find that participants are more likely to correctly *not* identify unnecessary actions than to correctly name indicated actions. For example, in the information arm the proportion of correct responses is 91.5 percent for unlisted and unnecessary actions as opposed to 28 percent for unlisted but indicated actions. This difference may arise from several factors: for instance, not naming a wrong action may be easier than naming a correct action or the two unlisted unnecessary actions—not performing an unnecessary cesarean section and not referring incorrectly—may be particularly salient.

6.2 Spillovers from incentives on unlisted actions and listed but unpaid actions

In addition to the direct effects on paid indicated and unnecessary actions, Table 4 also summarizes estimated indirect effects on unlisted actions or listed but unpaid actions. For unlisted actions, we find evidence consistent with positive spillovers on indicated actions: performance in the two incentive arms is 4.1 (rewards) and 3.8 (penalties) pp higher than in the information arm, in which

respondents named 28.3 percent of correct actions, on average. Table 1 shows that this effect is driven by two unlisted actions, measuring the mother's vital signs and the rate of descent of the fetal head. We find no evidence of spillovers for unnecessary actions. We also do not find spillovers for listed but unincentivized actions; performance on these actions is between 10.8 and 12.7 percent across the arms, which may be because incentivizing actions increases their salience relative to the listed-but-incentivized actions. In the context of the framework described above, we interpret positive spillovers from incentives as indicative of complementarities across actions.

6.3 Potential confounders

We examine two potential confounders of our estimated impacts: whether the facility's participation in the concurrent PFP trial led to a differential impact on participants' performance and whether participants' knowledge or skills of maternity care affects their performance on the experiment.

Table 5 presents interacted regressions of treatment assignment in our experiment with the larger trial's arms, control, DFF and PFP. Table A.6 presents results disaggregated by case. The results show that our estimated impacts of incentives are robust to the inclusion of assignment to the PFP trial and interaction terms. Those assigned to our rewards and penalties arms always perform better than those in the information arm. As in the larger impact evaluation by Kandpal et al. (2019), we find that participants in the DFF or PFP arms perform better than those in the matched control arm. The lowest-performing group are participants in the control arm of the larger trial who were assigned to the information arm in our study. Nonetheless, the robustness of the main estimated impacts suggests that prior exposure to PFP does not drive the responses to our task. Our ability to replicate the qualitative findings from the larger trial also bolsters our confidence in our measure of performance, which is different from what the larger impact evaluation used.

In Table A.7 we further examine whether awareness and understanding of the larger PFP trial confound responses to the experimental tasks. We create a binary measure for awareness using responses to a survey question of whether the respondents' health facility participates in the trial and a second a binary measure of (above-median) understanding based on questions about

how many indicators are incentivized in the larger trial's PFP arm. We find that awareness and understanding are associated with higher performance (as measured across responses), but this effect does not covary with our study arms. Thus, our results suggest that neither facility-level participation in the larger trial nor worker-level awareness that the facility participates in the PFP trial has a significant moderating effect on performance on the partograph task.

We examine knowledge as potential confounder in columns 3 and 4 of Table 5. Scoring above the median on the antenatal care vignette is positively correlated with performance. However, knowledge does not change the sign and statistical significance of our main estimates. The estimates for each of the five partograph cases in Table A.6 also generally suggest that participants with a higher knowledge score do not respond differently to the incentives than workers with a lower knowledge score. We find a similar pattern when examining how performance in the experiment correlates with performance when costly effort is involved, i.e. in adherence to protocol on screening for five common danger signs in pregnancy in actual patient-provider interactions. We examine three outcome measures related to such screening: whether the health worker screened the pregnant woman for any danger signs, all danger signs, and the total number of danger signs screened for. All three measures of real effort are correlated with performance in our experimental task (Table A.8) and accounting for them does not substantively affect our impact estimates. Together these findings suggest that knowledge does not moderate performance in our experimental task, but also that our task captures actual effort and attention rather than only knowledge or skill.

6.4 Benchmarking performance against no intervention

We can benchmark the performance of information against no intervention by comparing, in the information arm, performance on listed actions with those that are unlisted (Table 4). Subject to the caveats discussed in Section 5, performance is 28.3 percent on unlisted actions and 26.4 percent on listed actions, suggesting information alone had no effect in our experiment. If the positive spillovers we find in the incentive arms also apply to the information arm, then the level of performance on unlisted actions would be higher than in a pure control arm and our estimate of the effect of information would be a lower bound.

6.5 Price response

We can exploit variation in the price assigned to tasks to examine how performance changes with the amount of incentive. While we randomly assigned actions to be incentivized, we purposively assigned higher prices to more complex actions. For instance, correct referrals are priced at 300 Naira, while monitoring contractions is priced at 50 Naira. Performance on a given action thus reflects responses to both the actions's price and non-price characteristics. We can recover the price response by netting out the level of performance on each action in the information arm, where actions only differ in their non-price characteristics. Figure 2 plots the percentage point difference in the incentive arms relative to the information arm, for actions from the three complex cases that are listed and indicated. Going from zero price to the lowest price of 50 Naira increases effort by 7 percentage points and the impact does not increase further in the incentive amount.

Based on these estimates, we calculate price elasticities of effort between 0.08 and 0.50 for each of the paid and indicated actions (Table A.12). While these estimates may at first glance appear small, they are comparable to the range of wage elasticities estimated by Oettinger (1999) and Goldberg (2016).[9] These low wage elasticities are also reassuring about the validity of our task. Since participants do not actually need to perform the action they identified, one might expect them to respond to the price to a greater degree than when they would have to incur substantial effort costs, which would artificially inflate our elasticity estimates. That, despite such potential upward bias, our elasticity estimates are in line with previous estimates, emphasizes the validity of our task.

Finally, as Figure 2 suggests, we also find effort to respond more to a low price than to incremental increases in price. This tapering off suggests that the key role of the financial incentive is to signal the salience of the task. This is consistent with evidence from public finance and environmental economics on the interaction of salience and financial (dis)incentives (Chetty et al., 2009; Sexton, 2015). It also aligns with evidence that anti-poverty cash transfers to households act as nudges to increase the salience of the behavior on which the transfer is conditioned (Benhassine

[9]These estimates are significantly smaller than the experimental 1.12-1.25 wage elasticity estimated by (Fehr and Goette, 2007) for bike messengers in Zurich. Indeed, the authors note that their estimated elasticities are much larger than those generally reported in the literature. In this regard, our estimates are more aligned with the literature.

et al., 2015) and that larger transfers may not necessarily increase the behavioral response (Filmer and Schady, 2011).

6.6 Health impact

We can perform a rudimentary calculation of the potential health gain by translating the increase in adherence to labor and delivery protocols to mortality gains for newborns within the first 7 days of birth (early neonatal mortality), which is likely most malleable to the effort applied by the health worker. We do so using estimates of the impact of protocol adherence on neonatal mortality. Because there are additional benefits of better protocol adherence—to the mother and newborn—we likely under-estimate the health gains.

In a study of delivery care in health facilities in Uttar Pradesh, India, Semrau et al. (2020) estimate that each additional action (out of 10 actions) by the health care provider is associated with a 30 percent decrease in early neonatal mortality. We observe an 8 percent improvement in the rewards arm relative to the information arm, corresponding to 0.8 additional actions. Assuming linearity, this would imply a 24 percent reduction in early neonatal mortality among births in health facilities. Using the observed neonatal mortality of 33 per 1,000 deliveries in (Semrau et al., 2020), the 24 percent reduction translates into 8 averted early neonatal deaths per 1,000 facility-based deliveries. Of the approximately 7.6 million births in Nigeria in 2018, 39 percent, or roughly 3 million, occurred in health facilities (National Population Commission, 2019). Thus, the improvement in protocol adherence we observed impact would translate into 24,000 fewer neonatal deaths.

We can benchmark the size of this impact in two ways. The first is the economic benefit. The value of a statistical life in Nigeria is estimated to be USD 485,000 (Viscusi and Masterman, 2017) with a life expectancy of 55 years in 2018 (World Bank, 2019). Thus, 24,000 fewer neonatal deaths would translate into an annualized economic benefit of USD 212 million. We can also compare our estimated health gain to that of alternative policies. Okeke (2021) reports on a cluster randomized trial in Nigeria in which either qualified physicians or mid-level professionals are sent to primary care health facilities. They find that physicians produce significantly higher quality of

antenatal and delivery care, translating into an intent-to-treat impact of 6-8 fewer early neonatal deaths per 1,000 live births. Okeke (2021) estimates that more sustained contact with physicians over the course of the pregnancy translates into a mortality reduction of 9-13 deaths per thousand. The authors also note that this magnitude of improvement is equivalent to the entire mortality reduction Nigeria achieved between 1990 and 2017. Thus, our estimated impact of 8 neonatal deaths averted from improving health worker effort on the intensive margin is comparable to the short-run extensive margin gains from adding an additional physician to a primary health facility. Because only 39 percent of all births in Nigeria occur in health facilities, other approaches may have an even greater impact on neonatal mortality. For instance, Okeke and Abubakar (2020) estimate that a conditional cash transfer in Nigeria prevented up to 85,000 neonatal deaths nationally. Compared to our estimated impact of 24,000, this much larger impact of a cash transfer reflects the fact that most neonatal mortality risk is in fact for births outside of health facility settings.

We lack the cost data required for a full cost-effectiveness analysis of implementing this experimental PFP contract at scale. The incentives arms in our experimental task had outlays that were 5 percent higher those in the information arm. Actual PFP programs in LMICs have substantial administrative costs, including for training, verifying data, and executing payments (Fritsche et al., 2014). For example, in the concurrent PFP trial, about 36.5 percent of total program costs were for administration and operations rather than disbursements to health facilities (Zeng et al., 2021).

7 Discussion

PFP schemes have been implemented in advanced health systems, such as in the United States and United Kingdom (see for example, Doran et al., 2011), and are also increasingly prevalent in LMICs. PFP contracts can elicit effort through two economic channels: information and financial incentives. In addition, incentives may elicit different levels of effort depending on whether they are cast as rewards or penalties. In this paper, we report on a framed field experiment designed to examine the direct effect of adding rewards or penalties to information on actions that are incentivized in a PFP contract as well as any indirect effects on actions that are merely listed or unlisted. We randomized maternity care workers in Nigeria into three arms (information only, rewards, and penalties) and

listed or incentivized a subset actions in labor and delivery care that are either clinically indicated or unnecessary. We observed the universe of clinical actions that a health worker could take in any situation and measured performance as adherence to clinical protocols.

Our findings indicate that incentives matter above and beyond information: the two incentive arms significantly outperform the information arm. However, the type of incentive does not seem to make a difference: rewards and penalties perform similarly well compared to information alone. Specifically, we estimate that both rewards and penalties increase overall performance by approximately 4 pp (about 8 percent) relative to the information arm. This effect is driven by an increase of 8 pp (about 20 percent) in incentivized actions. We also find evidence of positive spillovers from rewards and penalties on unlisted (and unincentivized) actions that are clinically indicated. Performance on these actions is 4 pp (about 14 percent) higher in the two incentivized arms than in the information arm. Further, while our experimental task precluded the inclusion of a pure control arm—the partographs themselves could have conveyed information—we obtain a benchmark effect of providing information by comparing performance on unlisted and listed actions in the information arm. We do not find that information alone increases performance. In the context of our conceptual framework, we interpret the direct effects as evidence that incentives are a critical component of PFP contracts, the positive spillovers as indication that actions are complements in production, and the similar effects of the rewards and penalties as an indication that participants are loss neutral. Our finding that effort changes mostly when the price jumps from zero to a positive price suggests that incentives may serve as a signal.

While our study design allows us to isolate the incentive effects, there are important caveats to the external validity of our results. While our design and the experimental task—as well as observed behavior—are realistic, the participant responses may not reflect costly real effort or trade-offs in clinical practice. Nonetheless, we establish that performance on our task is significantly correlated with both health worker knowledge and, for a randomly-selected subset of our sample, performance in actual patient-provider interactions. Similarly, because the task is hypothetical, we do not capture possible effects of altruism, which could interact with the interventions if, say, offering financial incentives erodes altruism (Lohmann et al., 2016). Third, our study examines responses to PFP among current health workers and cannot speak to the effect of PFP on workforce

composition. Recent evidence shows that PFP can have substantial compositional effects on the teacher and health provider workforce through differential recruitment, which may lead to lower or higher performance (Deserranno, 2019; Leaver et al., forthcoming) and, possibly, to different responses to PFP. While our paper cannot speak to the effect of PFP on this margin, public health systems in LMICs often have rigid recruitment and career progression regulations that may limit effect on workforce composition, at least in the short run (Araujo and Maeda, 2013). Finally, while our simple PFP contract—which provides high-powered incentives directly to the worker—improves health worker performance, these results may not translate to "real-world" PFP schemes, which are more complex and may be harder to understand. Our participants were immediately and directly paid at the end of the experiment, while performance-based bonuses in the concurrent trial were calculated at the facility level, transferred to the facility, and only then apportioned among staff.[10] The impact evaluation of the concurrent Nigeria PFP trial found no additional impact of PFP over decentralized facility financing, despite substantial payouts (Kandpal et al., 2019). The evaluation also reported that three-quarters of the health workers in the facilities in the PFP arm could not correctly identify the actions that would increase their performance bonus and over half had not even heard of the PFP trial.

Nonetheless, our experiment provides new evidence characterizing the information and incentive channels of PFP contracts as well as the spillovers from incomplete contracts. We also find loss framing not to matter for either direct or spillover effects. Disentangling these mechanisms is important for our understanding of contracting arrangements to resolve principal-agent problems and provides guidance on how to empirically examine the information and incentive channels (Prendergast, 1999). Taken at face value, our results imply that direct, high-powered PFP incentives would outperform health worker interventions that provide the same information but without incentives. In particular, the health gains implied by our crude calculations suggest that PFP is about as effective as increasing the health workforce by one qualified physician per primary health facility in Nigeria. Contracts with rewards appear to generate the same performance gain as penalties and may be preferable for administrative ease and political acceptability. However, the observed price response suggests incentives may chiefly function as a signal of importance and that small

[10] Teacher unions have imposed this sort of school-level structure on the design of PFP contracts in education. (Fryer et al., 2012)

incentives may be sufficient for this purpose. For this reason, PFP contracts could be made more cost-effective by making smaller payments that merely signal the importance of the task.

References

E. Araujo and A. Maeda. How to recruit and retain health workers in rural and remote areas in developing countries: A guidance note. HNP Discussion Paper, World Bank, 2013.

K. J. Arrow. Uncertainty and the welfare economics of medical care. *American Economic Review*, 53(5):941–973, 1963.

N. Ashraf, O. Bandiera, and B. K. Jack. No margin, no mission? a field experiment on incentives for public service delivery. *Journal of Public Economics*, 120:1–17, 2014.

P. Basinga, P. J. Gertler, A. Binagwaho, A. L. Soucat, J. Sturdy, and C. M. Vermeersch. Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: An impact evaluation. *The Lancet*, 377(9775):1421–1428, 2011.

N. Benhassine, F. Devoto, E. Duflo, P. Dupas, and V. Pouliquen. Turning a shove into a nudge? a "labeled cash transfer" for education. *American Economic Journal: Economic Policy*, 7(3), 2015.

J. M. Brock, A. Lange, and K. L. Leonard. Giving and promising gifts: Experimental evidence on reciprocity from the field. *Journal of Health Economics*, 58:188–201, 2018.

E. Bulte, J. A. List, and D. Van Soest. Toward an understanding of the welfare effects of nudges: Evidence from a field experiment in the workplace. *The Economic Journal*, 130(632):2329–2353, 2020.

E. Bulte, J. A. List, and D. van Soest. Incentive spillovers in the workplace: Evidence from two field experiments. *Journal of Economic Behavior & Organization*, 184:137–149, 2021.

S. Campbell, D. Reeves, E. Kontopantelis, E. Middleton, B. Sibbald, and M. Roland. Quality of primary care in england with the introduction of pay for performance. *New England Journal of Medicine*, 357(2):181–190, 2007.

M. A. Cattaneo, C. Oggenfuss, and S. C. Wolter. The more, the better? the impact of instructional time on student performance. *Education economics*, 25(5):433–445, 2017.

P. A. Celhay, P. J. Gertler, P. Giovagnoli, and C. Vermeersch. Long-run effects of temporary incentives on medical care productivity. *American Economic Journal: Applied Economics*, 11 (3):92–127, 2019.

R. Chetty, A. Looney, and K. Kroft. Salience and taxation: Theory and evidence. *American Economic Review*, 99(4), 2009.

J. Das and P. J. Gertler. Variations in practice quality in five low-income countries: A conceptual overview. *Health Affairs*, 26(Suppl2):w296–w309, 2007.

J. Das, J. Hammer, and K. Leonard. The quality of medical advice in low-income countries. *Journal of Economic Perspectives*, 22(2):93–114, 2008.

J. Das, A. Holla, V. Das, M. Mohanan, D. Tabak, and B. Chan. In urban and rural india, a standardized patient study showed low levels of provider training and huge quality gaps. *Health affairs*, 31(12):2774–2784, 2012.

J. Das, A. Holla, A. Mohpal, and K. Muralidharan. Quality and accountability in health care delivery: Audit-study evidence from primary care in india. *American Economic Review*, 106 (12):3765–3799, 2016.

J. de Quidt. Your loss is my gain: a recruitment experiment with framed incentives. *Journal of the European Economic Association*, 16(2):522–559, 2018.

J. de Quidt, F. Fallucchi, F. Kölle, D. Nosenzo, and S. Quercia. Bonus versus penalty: How robust are the effects of contract framing? *Journal of the Economic Science Association*, 3(2):174–182, 2017.

S. DellaVigna and D. Pope. What motivates effort? evidence and expert forecasts. *Review of Economic Studies*, 85(2):1029–1069, 2018.

E. Deserranno. Financial incentives as signals: Experimental evidence from the recruitment of village promoters in Uganda. *American Economic Journal: Applied Economics*, 11(1):277–317, 2019.

K. Diaconu, J. Falconer, A. V. Verbel Facuseh, A. Fretheim, and S. Witter. Paying for performance to improve the delivery of health interventions in low- and middle-income countries. *Cochrane Database of Systematic Reviews*, (12):CD007899, 2020.

T. Doran, E. Kontopantelis, J. M. Valderas, S. Campbell, M. Roland, C. Salisbury, and D. Reeves. Effect of financial incentives on incentivised and non-incentivised clinical activities: Longitudinal analysis of data from the UK Quality and Outcomes Framework. *British Medical Journal*, 342: d3590, 2011.

E. Duflo, R. Hanna, and S. P. Ryan. Incentives work: Getting teachers to come to school. *American Economic Review*, 102(4):1241–78, 2012.

E. Fehr and L. Goette. Do workers work more if wages are high? evidence from a randomized field experiment. *American Economic Review*, 97(1):298–317, 2007.

D. Filmer and N. Schady. Does more cash in conditional cash transfer programs always lead to larger impacts on school attendance? *Journal of Development Economics*, 96(1):150–157, 2011.

F. Finan, B. A. Olken, and R. Pande. The personnel economics of the state. NBER Working Paper w21825, National Bureau of Economic Research, 2015.

National Population Commission. Nigeria demographic and health survey 2018 - final report. Technical report, 2019.

G. B. Fritsche, R. Soeters, and B. Meessen. *Performance-based financing toolkit*. The World Bank, 2014.

R. G. Fryer, S. D. Levitt, J. List, and S. Sadoff. Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. Technical report, National Bureau of Economic Research, 2012.

V. Gauri, J. C. Jamison, N. Mazar, O. Ozier, S. Raha, and K. Saleh. Motivating bureaucrats through social recognition: evidence from simultaneous field experiments. Technical report, The World Bank, 2018.

J. Goldberg. Kwacha gonna do? experimental evidence about labor supply in rural malawi. *American Economic Journal: Applied Economics*, 8(1):129–49, 2016.

A. Gupta. Impacts of performance pay for hospitals: The readmissions reduction program. *American Economic Review*, 111(4):1241–83, 2021.

G. W. Harrison and J. A. List. Field experiments. *Journal of Economic Literature*, 42(4), December 2004.

B. Holmstrom and P. Milgrom. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, 7:24, 1991.

T. Hossain and J. A. List. The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science*, 58(12):2151–2167, 2012.

A. Imas, S. Sadoff, and A. Samek. Do people anticipate loss aversion? *Management Science*, 63 (5):1271–1284, 2017.

D. Kahneman and A. Tversky. Choices, values, and frames. *American Psychologist*, 39(4):341–350, 1984.

D. Kahneman, J. L. Knetsch, and R. H. Thaler. Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, 5(1):193–206, 1991.

E. Kandpal, B. P. Loevinsohn, C. M. Vermeersch, E. Pradhan, M. Khanna, M. K. Conlon, and W. Zeng. Impact evaluation of Nigeria State Health Investment Project. Technical report, The World Bank, 2019.

J. T. Kolstad. Information and quality when motivation is intrinsic: Evidence from surgeon report cards. *American Economic Review*, 103(7):2875–2910, 2013.

V. Lavy. Expanding school resources and increasing time on task: Effects on students' academic and non-cognitive outcomes. *Journal of the European Economic Association*, 2016.

E. P. Lazear. Performance pay and productivity. *American Economic Review*, 90(5):1346–1361, December 2000.

C. Leaver, O. W. Ozier, P. M. Serneels, and A. Zeitlin. Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from Rwandan primary schools. *American Economic Review*, forthcoming.

K. L. Leonard and M. C. Masatu. Professionalism and the know-do gap: Exploring intrinsic motivation among health workers in Tanzania. *Health Economics*, 19(12):1461–1477, 2010.

K. L. Leonard and M. C. Masatu. Changing health care provider performance through measurement. *Social Science & Medicine*, 181:54–65, 2017.

K. L. Leonard, M. C. Masatu, and A. Vialou. Getting doctors to do their best the roles of ability and motivation in health care quality. *Journal of Human Resources*, 42(3):682–700, 2007.

J. Lohmann, N. Houlfort, and M. De Allegri. Crowding out or no crowding out? A self-determination theory approach to health worker motivation in performance-based financing. *Social Science & Medicine*, 169:1–8, 2016.

C. Lopez, A. Sautmann, and S. Schaner. Does patient demand contribute to the overuse of prescription drugs? *American Economic Journal: Applied Economics*, forthcoming.

T. G. McGuire. Physician agency. In A. J. Culyer and J. Newhouse, editors, *Handbook of Health Economics*, volume 1A, page 461–536. 2000.

A. Mendelson, K. Kondo, C. Damberg, A. Low, M. Motúapuaka, M. Freeman, M. O'neil, R. Relevo, and D. Kansagara. The effects of pay-for-performance programs on health, health care use, and processes of care: A systematic review. *Annals of Internal Medicine*, 166(5):341–353, 2017.

G. Miller and K. S. Babiarz. Pay-for-performance incentives in low- and middle-income country health programs. Working Paper 18932, National Bureau of Economic Research, April 2013.

K. J. Mullen, R. G. Frank, and M. B. Rosenthal. Can you get what you pay for? pay-for-performance and the quality of healthcare providers. *The RAND Journal of Economics*, 41(1): 64–91, Mar 2010.

G. S. Oettinger. An empirical analysis of the daily labor supply of stadium vendors. *Journal of political Economy*, 107(2):360–392, 1999.

E. N. Okeke. When a doctor falls from the sky: The impact of easing physician supply constraints on mortality. Technical report, Mimeo, 2021.

E. N. Okeke and I. S. Abubakar. Healthcare at the beginning of life and child survival: Evidence from a cash transfer experiment in nigeria. *Journal of Development Economics*, 143:102426, 2020.

L. Pierce, A. Rees-Jones, and C. Blank. The negative consequences of loss-framed performance incentives. Technical report, National Bureau of Economic Research, 2020.

C. Prendergast. The provision of incentives in firms. *Journal of Economic Literature*, 37(1):7–63, Mar. 1999.

S. G. Rivkin and J. C. Schiman. Instruction time, classroom quality, and academic achievement. *The Economic Journal*, 125(588):F425–F448, 2015.

J. Rothstein. Teacher quality policy when supply matters. *American Economic Review*, 105(1): 100–130, January 2015.

A. K. Rowe, D. De Savigny, C. F. Lanata, and C. G. Victora. How can we achieve and maintain high-quality performance of health workers in low-resource settings? *The Lancet*, 366(9490): 1026–1035, 2005.

K. E. Semrau, K. A. Miller, S. Lipsitz, J. Fisher-Bowman, A. Karlage, B. A. Neville, M. Krasne, J. Gass, A. Jurczak, V. Pratap Singh, S. Singh, M. Marx Delaney, L. R. Hirschhorn, B. Kodkany, V. Kumar, and A. A. Gawande. Does adherence to evidence-based practices during childbirth prevent perinatal mortality? a post-hoc analysis of 3,274 births in uttar pradesh, india. *BMJ Global Health*, 5(9), 2020.

S. Sexton. Automatic bill payment and salience effects: Evidence from electricity consumption. *Review of Economics and Statistics*, 97(2):229–241, 2015.

T. Sherry. A note on the comparative statics of pay-for-performance in health care. *Health Economics*, 25(5):637–644, 2016.

T. B. Sherry, S. Bauhoff, and M. Mohanan. Multitasking and heterogeneous treatment effects in pay-for-performance in health care: Evidence from Rwanda. *American Journal of Health Economics*, 3(2):192–226, 2017.

J. Villar, H. Ba'aqeel, G. Piaggio, P. Lumbiganon, J. M. Belizán, U. Farnot, Y. Al-Mazrou, G. Carroli, A. Pinol, A. Donner, et al. Who antenatal care randomised trial for the evaluation of a new model of routine antenatal care. *The Lancet*, 357(9268):1551–1564, 2001.

W. K. Viscusi and C. J. Masterman. Income elasticities and global values of a statistical life. *Journal of Benefit-Cost Analysis*, 8(2):226–250, 2017.

White Ribbon Alliance. Respectful maternity care: A Nigeria-focused health workers' training guide. Technical report, Futures Group, Health Policy Project, 2015.

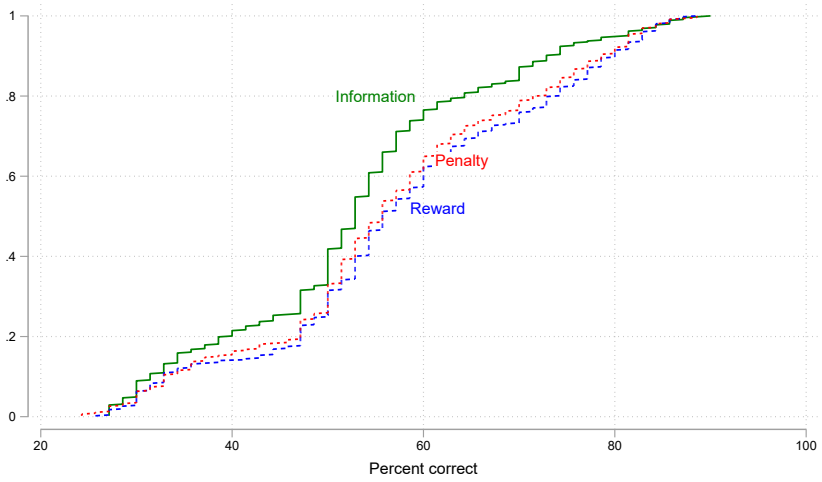
WHO. WHO recommendations for augmentation of labour. Technical report, World Health Organization, 2014.

World Bank. Life expectancy, 2019. data retrieved from World Development Indicators, https://data.worldbank.org/indicator/SP.DYN.LE00.IN?locations=NG.

W. Zeng, E. Pradhan, and M. Khanna. Cost-effectiveness analysis of the decentralized facility financing and performance-based financing program in nigeria. Technical report, Mimeo, 2021.

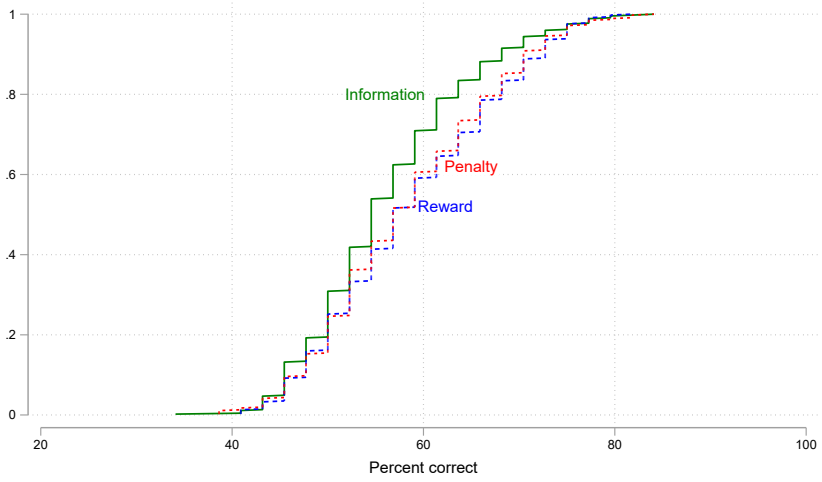
Figures and Tables

Figure 1: Empirical cumulative distributions



P-values for two-sample Kolmogorov-Smirnov test for equality of distribution:
Information < Reward p=0.00; Information < Penalty p=0.00; Reward < Penalty p=0.96 and Reward > Penalty p=0.28.

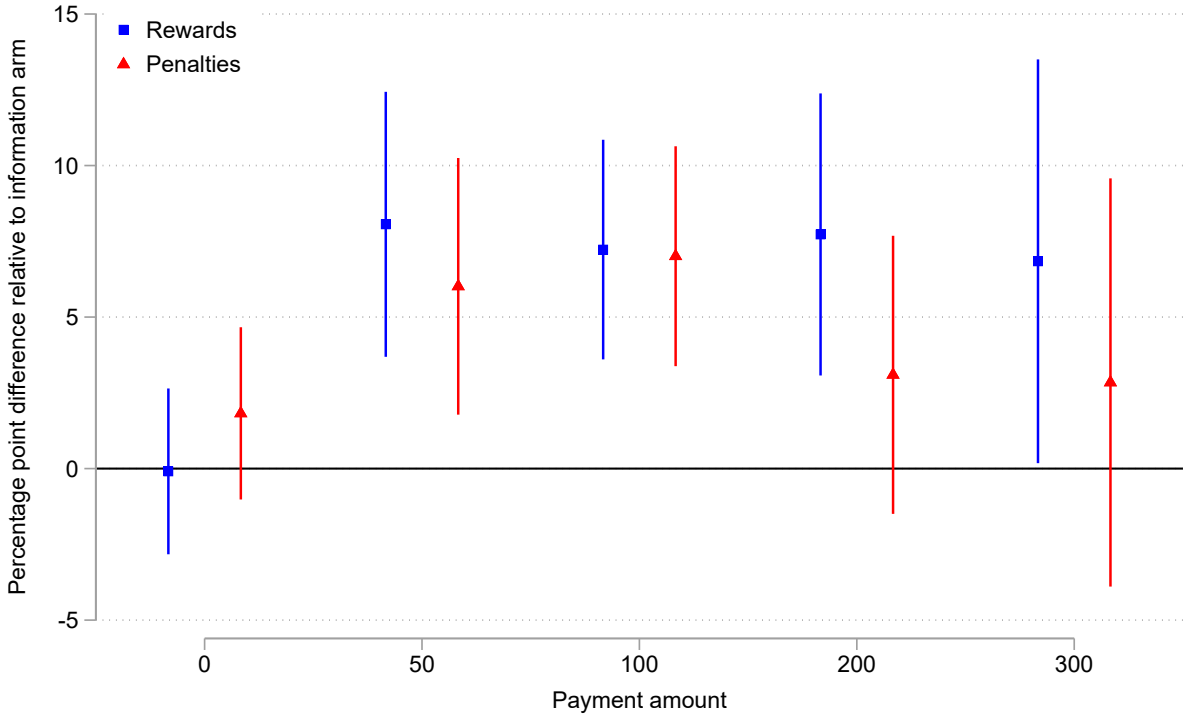
(a) Across cases: Cases weighted equally



P-values for two-sample Kolmogorov-Smirnov test for equality of distribution:
Information < Reward p=0.00; Information < Penalty p=0.00; Reward < Penalty p=0.97 and Reward > Penalty p=0.67.

(b) Across responses: Responses weighted equally

Figure 2: Percent correct by incentive amount



Based on indicated and listed actions in the three complex cases. For the complete regression results, see Table A.9.

Table 1: Prices and performance by type of actions

		Percent correct			
	Incentive (Naira)	All arms	Information	Reward	Penalty
Listed and paid					
Refer when necessary	300	75	72	78	76
Do not refer when unnecessary	200	70	66	74	71
Palpate the uterus	100	63	64	63	63
Monitor contractions	50	44	39	47	47
Monitor fetal heart rate	100	43	36	46	47
Listed but unpaid					
Monitor color and consistency of liquor		17	16	16	18
Record fluids/drugs administered		6	6	6	7
Unlisted					
Administer magnesium sulfate		96	97	97	96
Measure urine and test for protein/glucose		94	94	95	94
Augment labor		91	91	92	91
Repeat cervical exam now		82	84	81	81
Administer antibiotics		65	65	66	65
Prepare for imminent delivery		53	53	54	51
Measure rate of descent of fetal head		48	46	48	49
Measure mother's vital signs		37	32	39	38

Based on the three complex cases.

Table 2: Summary statistics and balance across experiment arms (percent)

Variable	(1) Information Mean/SE	(2) Reward Mean/SE	(3) Penalty Mean/SE	T-test Difference (1)-(2)	(1)-(3)	(2)-(3)
Outcomes†						
Across cases	52.90	57.76	56.64	-4.86***	-3.73***	1.12
	(0.69)	(0.72)	(0.71)			
Across responses	56.20	58.72	58.37	-2.52***	-2.17***	0.35
	(0.40)	(0.44)	(0.44)			
Total payout (Naira)	1,750.00	1,841.90	1,818.63	-91.90***	-68.63***	23.28
	(0.00)	(14.70)	(15.18)			
Covariates						
Male	27.52	22.10	20.70	5.42*	6.82**	1.40
	(2.11)	(1.94)	(1.89)			
Age $\geq$ median	50.11	54.49	53.81	-4.37	-3.70	0.67
	(2.37)	(2.33)	(2.33)			
Doctor or nurse	7.38	10.94	10.24	-3.56*	-2.86	0.70
	(1.24)	(1.46)	(1.42)			
Years since qualified $\geq$ median	49.89	49.89	52.07	-0.00	-2.18	-2.18
	(2.37)	(2.34)	(2.33)			
Knowledge $\geq$ median	52.57	52.08	51.20	0.49	1.37	0.88
	(2.36)	(2.34)	(2.34)			
NSHIP pilot status						
PFP	44.74	42.67	42.70	2.07	2.04	-0.03
	(2.35)	(2.32)	(2.31)			
DFF	42.06	45.08	42.92	-3.02	-0.86	2.16
	(2.34)	(2.33)	(2.31)			
Control	13.20	12.25	14.38	0.95	-1.18	-2.13
	(1.60)	(1.54)	(1.64)			
N	447	457	459			

Notes: †Across cases weighs each case equally; across responses weighs each response equally. The median for knowledge is 51.52 percent. The value displayed for t-tests are the differences in the means across the groups. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 3: Overall performance (percent correct)

	All cases		By case				
	(1) Across cases	(2) Across responses	(3) Simple 1	(4) Simple 2	(5) Complex 1	(6) Complex 2	(7) Complex 3
Reward	4.32***	2.04***	5.34*	11.03***	1.30*	1.25**	2.71***
	(0.97)	(0.45)	(3.02)	(3.08)	(0.75)	(0.51)	(0.84)
Penalty	3.36***	1.88***	2.56	9.19***	1.31	1.19**	2.55***
	(1.00)	(0.45)	(3.34)	(3.34)	(0.84)	(0.54)	(0.79)
Constant (Information)	53.21***	56.46***	68.58***	26.81***	62.33***	51.97***	56.33***
	(0.58)	(0.26)	(1.86)	(1.91)	(0.47)	(0.31)	(0.48)
Facility fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
P-value Penalty v Reward	0.292	0.700	0.374	0.544	0.988	0.917	0.836
N respondents	1,363	1,363	1,363	1,363	1,363	1,363	1,363
R-squared (overall)	0.018	0.015	0.003	0.011	0.005	0.006	0.010

* p<0.10, ** p<0.05, *** p<0.01. OLS models with facility fixed effects and robust standard errors.

Table 4: Percent correct by arm and type of action

	Indicated actions				Contra-indicated actions	
	Unlisted	Listed	Paid	Listed or paid	Unlisted	Paid
Level (%)						
Information	28.3	10.9	38.1	26.4	91.5	79.4
Reward	32.4	10.8	45.5	30.6	90.9	82.3
Penalty	32.1	12.7	46.0	31.7	89.9	80.6
Difference (% points)						
Reward - Information	4.1	-0.1	7.4	4.2	-0.6	2.9
p-value	0.01	0.95	0.00	0.00	0.55	0.02
Penalty - Information	3.8	1.8	7.9	5.3	-1.7	1.1
p-value	0.02	0.21	0.00	0.00	0.13	0.36
Penalty - Reward	-0.4	1.9	0.5	1.1	-1.0	-1.7
p-value	0.83	0.19	0.77	0.45	0.33	0.16

Differences from unadjusted OLS models; s.e. clustered at worker level. The full output is reported in Table A.5. Based on the three complex cases.

Table 5: Interaction with status in concurrent trial of health facility financing modalities and knowledge on vignette

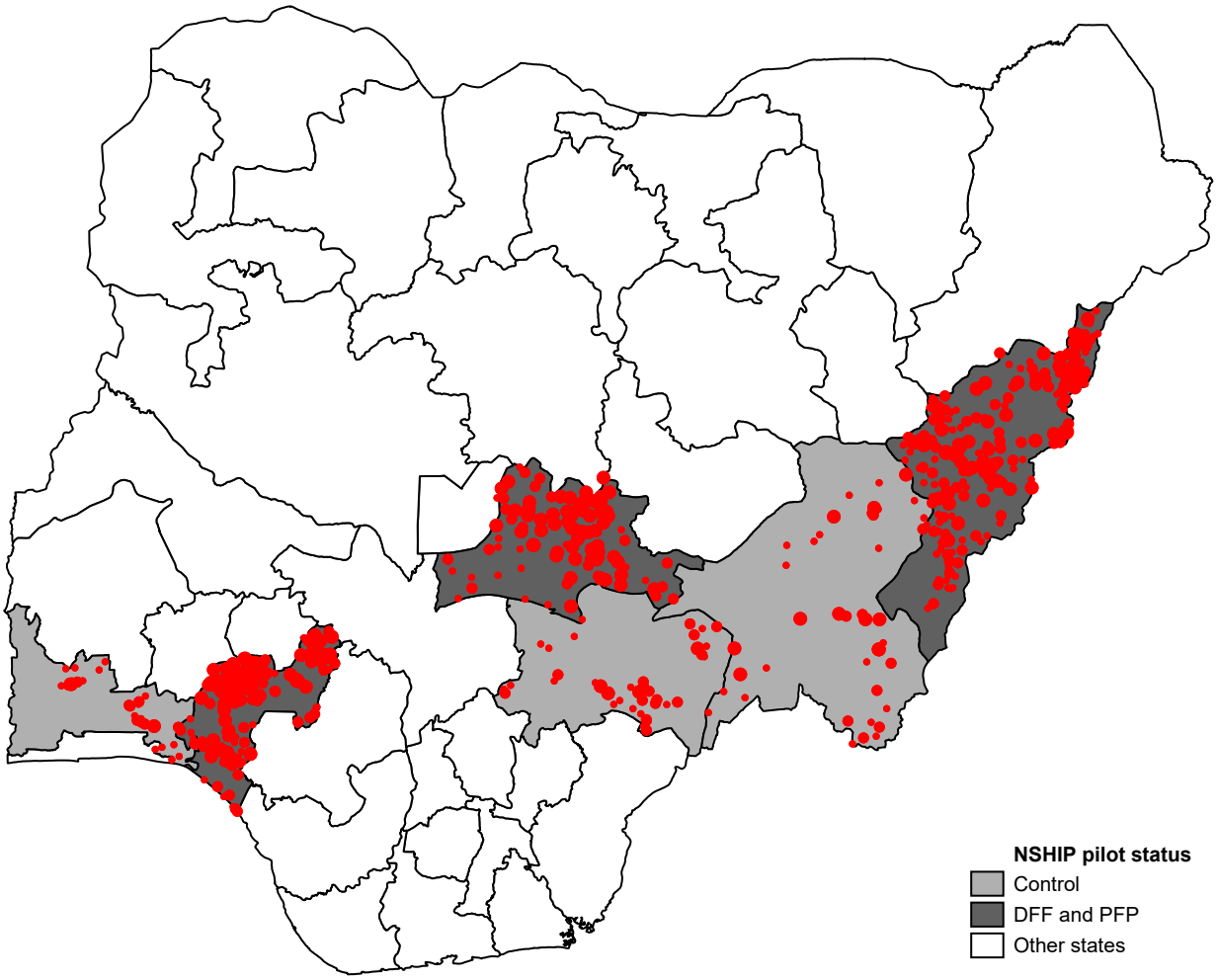
	Larger trial		Knowledge	
	(1) Across cases	(2) Across responses	(3) Across cases	(4) Across responses
Reward	9.82***	2.73**	4.09***	3.08***
	(2.71)	(1.18)	(1.48)	(0.66)
Penalty	4.34*	1.40	3.04*	2.78***
	(2.54)	(1.13)	(1.55)	(0.67)
DFF	5.04**	3.62***		
	(2.05)	(0.99)		
PFP	6.07***	5.04***		
	(2.10)	(1.00)		
Reward × DFF	-5.96*	-0.55		
	(3.06)	(1.48)		
Reward × PFP	-5.40*	0.09		
	(3.15)	(1.52)		
Penalty × DFF	-1.23	0.66		
	(2.91)	(1.44)		
Penalty × PFP	0.00	1.31		
	(2.99)	(1.48)		
Knowledge ≥ median			1.92	3.10***
			(2.05)	(0.87)
Reward × Knowledge ≥ median			0.49	-1.94**
			(2.06)	(0.97)
Penalty × Knowledge ≥ median			0.61	-1.73*
			(2.08)	(0.94)
Constant (Information × Control)	48.06***	52.43***	52.20***	54.84***
	(1.80)	(0.77)	(1.26)	(0.52)
Facility fixed effects	No	No	Yes	Yes
P-values from tests of coefficients				
Control: Penalty v Reward	0.043	0.277		
DFF: Penalty v Reward	0.124	0.425		
PFP: Penalty v Reward	0.089	0.440		
Penalty v Reward			0.447	0.666
N respondents	1,363	1,363	1,363	1,363

* p<0.10, ** p<0.05, *** p<0.01. Unadjusted OLS models; s.e. clustered at facility level. Estimates by partograph case are available in Table A.6.

Appendices

A Additional results

Figure A.1: Location of study clinics



The size of the marker is proportional to number of respondents (range 1–4). Trial status refers to the concurrent cluster-randomized trial. The intervention states are Adamawa, Nasarawa, and Ondo; the control states are Taraba, Benue, and Ogun.

Table A.1: Scoring scheme for each possible action in the complex tasks

	Complex 1	Complex 2	Complex 3
Listed and paid			
Refer when necessary	Unnecessary	Indicated	Unnecessary
Do not refer when unnecessary	Indicated	Unnecessary	Indicated
Palpate the uterus	Ambiguous	Indicated	Ambiguous
Monitor contractions	Indicated	Indicated	Indicated
Monitor fetal heart rate	Indicated	Indicated	Indicated
Listed but unpaid			
Monitor color and consistency of liquor	Indicated	Indicated	Indicated
Record fluids/drugs administered	Indicated	Indicated	Indicated
Unlisted			
Administer magnesium sulfate	Unnecessary	Unnecessary	Unnecessary
Measure urine and test for protein/glucose	Unnecessary	Unnecessary	Ambiguous
Augment labor	Unnecessary	Unnecessary	Unnecessary
Repeat cervical exam now	Unnecessary	Unnecessary	Unnecessary
Administer antibiotics	Unnecessary	Indicated	Unnecessary
Prepare for imminent delivery	Indicated	Unnecessary	Indicated
Measure rate of descent of fetal head	Indicated	Ambiguous	Indicated
Measure mother's vital signs	Indicated	Indicated	Indicated

Table A.2: Percent correct by action and case

	Overall	Complex 1	Complex 2	Complex 3
Listed and paid				
Refer when necessary	75		75	
Do not refer when unnecessary	70	79		61
Palpate the uterus	63	91	8	92
Monitor contractions	44	58	32	43
Monitor fetal heart rate	43	53	35	42
Listed but unpaid				
Monitor color and consistency of liquor	17	21	12	16
Record fluids/drugs administered	6	8	6	6
Unlisted				
Administer magnesium sulfate	96	96	96	97
Measure urine and test for protein/glucose	94	93	95	94
Augment labor	91	93	94	87
Repeat cervical exam now	82	78	86	84
Administer antibiotics	65	96	4	96
Prepare for imminent delivery	53	35	90	33
Measure rate of descent of fetal head	48	38	77	28
Measure mother's vital signs	37	45	30	34

Correct captures actions that participants named and are clinically indicated as well as actions that were not named and are unnecessary.

Table A.3: Interview length

	(1)	(2)
Reward	54.84	
	(41.80)	
Penalty	87.16**	
	(42.59)	
Incentives (Reward or Penalty)		70.91*
		(36.44)
Constant (Information)	626.78***	626.99***
	(24.64)	(24.65)
Facility fixed effects	Yes	Yes
P-value Penalty v Reward	0.449	
N respondents	1,178	1,178
R-squared (overall)	0.001	0.000

* p<0.10, ** p<0.05, *** p<0.01. OLS models with facility fixed effects and robust standard errors. Length of the full worker interview in minutes.

Table A.4: Overall performance (z-scores)

	All cases		By case				
	(1) Across cases	(2) Across responses	(3) Simple 1	(4) Simple 2	(5) Complex 1	(6) Complex 2	(7) Complex 3
Reward	0.28***	0.22***	0.12*	0.23***	0.10*	0.15**	0.20***
	(0.06)	(0.05)	(0.07)	(0.07)	(0.06)	(0.06)	(0.06)
Penalty	0.22***	0.21***	0.06	0.19***	0.10	0.14**	0.19***
	(0.07)	(0.05)	(0.07)	(0.07)	(0.06)	(0.06)	(0.06)
Constant (Information)	-0.17***	-0.14***	-0.06	-0.14***	-0.07*	-0.10***	-0.13***
	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)
Facility fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
P-value Penalty v Reward	0.292	0.700	0.374	0.544	0.988	0.917	0.836
N respondents	1,363	1,363	1,363	1,363	1,363	1,363	1,363
R-squared (overall)	0.018	0.015	0.003	0.011	0.005	0.006	0.010

* p<0.10, ** p<0.05, *** p<0.01. OLS models with facility fixed effects and robust standard errors.

Table A.5: Percent correct by arm for different actions

	Indicated actions								Counter-indicated actions			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Unlisted	Unlisted	Listed	Listed	Paid	Paid	Listed or Paid	Listed or Paid	Unlisted	Unlisted	Paid	Paid
Reward	4.14**	4.74***	-0.09	0.51	7.40***	7.89***	4.19***	4.73***	-0.62	-1.01	2.86**	2.66**
	(1.67)	(1.56)	(1.39)	(1.34)	(1.80)	(1.70)	(1.43)	(1.35)	(1.05)	(0.99)	(1.26)	(1.24)
Penalty	3.78**	4.47***	1.82	2.45*	7.94***	8.48***	5.32***	5.90***	-1.67	-2.15**	1.14	0.98
	(1.68)	(1.56)	(1.45)	(1.39)	(1.81)	(1.72)	(1.47)	(1.39)	(1.09)	(1.03)	(1.25)	(1.24)
Constant (Information)	28.27***	31.55***	10.85***	6.90***	38.09***	45.42***	26.41***	25.96***	91.54***	98.14***	79.42***	87.58***
	(1.17)	(1.54)	(0.99)	(1.26)	(1.23)	(1.88)	(0.99)	(1.37)	(0.75)	(0.88)	(0.91)	(1.20)
Worker covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Case FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N actions	10,904	10,904	8,178	8,178	10,904	10,904	19,082	19,082	21,808	21,808	5,452	5,452
R-squared	0.002	0.076	0.001	0.039	0.005	0.059	0.002	0.033	0.001	0.028	0.001	0.020

* p<0.10, ** p<0.05, *** p<0.01. OLS models; s.e. clustered at worker level. Actions from the three complex cases.

Table A.6: Overall performance (percent correct) for moderators

	All cases		By case				
	(1) Across cases	(2) Across responses	(3) Simple 1	(4) Simple 2	(5) Complex 1	(6) Complex 2	(7) Complex 3
A. Interaction with status in larger trial							
Reward	9.82***	2.73**	15.50*	28.15***	3.05	-0.89	3.29*
	(2.71)	(1.18)	(8.87)	(8.78)	(1.86)	(1.39)	(1.93)
Penalty	4.34*	1.40	4.67	13.97*	2.43	-0.26	0.88
	(2.54)	(1.13)	(8.86)	(8.29)	(1.80)	(1.33)	(1.77)
DFF	5.04**	3.62***	15.34**	-0.42	7.82***	-1.85	4.33***
	(2.05)	(0.99)	(7.28)	(6.51)	(1.63)	(1.15)	(1.59)
PFP	6.07***	5.04***	13.07*	2.58	8.94***	0.47	5.31***
	(2.10)	(1.00)	(7.27)	(6.52)	(1.61)	(1.13)	(1.65)
Reward × DFF	-5.96*	-0.55	-12.50	-17.71*	-2.72	3.20**	-0.06
	(3.06)	(1.48)	(9.95)	(9.91)	(2.31)	(1.61)	(2.33)
Reward × PFP	-5.40*	0.09	-9.11	-20.25**	0.01	2.47	-0.10
	(3.15)	(1.52)	(9.95)	(9.95)	(2.29)	(1.63)	(2.44)
Penalty × DFF	-1.23	0.66	-4.88	-3.94	-1.01	1.90	1.81
	(2.91)	(1.44)	(10.00)	(9.51)	(2.29)	(1.58)	(2.20)
Penalty × PFP	0.00	1.31	0.82	-5.24	0.36	1.70	2.37
	(2.99)	(1.48)	(9.95)	(9.53)	(2.28)	(1.58)	(2.30)
Constant (Information)	48.06***	52.43***	55.93***	25.42***	54.60***	52.42***	51.94***
	(1.80)	(0.77)	(6.48)	(5.69)	(1.27)	(0.98)	(1.29)
Facility fixed effects	No	No	No	No	No	No	No
P-values from tests of coefficients							
Control: Penalty v Reward	0.043	0.277	0.206	0.116	0.740	0.634	0.197
DFF: Penalty v Reward	0.124	0.425	0.430	0.177	0.456	0.407	0.411
PFP: Penalty v Reward	0.089	0.440	0.302	0.143	0.879	0.635	0.308
N respondents	1,363	1,363	1,363	1,363	1,363	1,363	1,363
B. Interaction with knowledge							
Reward	4.09***	3.08***	0.91	10.68**	3.42***	1.77**	3.66***
	(1.48)	(0.66)	(4.89)	(4.35)	(1.14)	(0.69)	(1.16)
Penalty	3.04*	2.78***	-0.90	7.85	4.23***	1.18	2.81**
	(1.55)	(0.67)	(5.35)	(5.12)	(1.14)	(0.76)	(1.16)
Knowledge ≥ median	1.92	3.10***	1.22	-1.34	6.16***	1.02	2.57
	(2.05)	(0.87)	(6.48)	(6.40)	(1.52)	(0.98)	(1.58)
Reward × Knowledge ≥ median	0.49	-1.94**	8.57	0.67	-4.00**	-0.98	-1.79
	(2.06)	(0.97)	(6.41)	(6.74)	(1.64)	(1.07)	(1.87)
Penalty × Knowledge ≥ median	0.61	-1.73*	6.61	2.53	-5.57***	0.00	-0.53
	(2.08)	(0.94)	(7.18)	(7.03)	(1.74)	(1.19)	(1.66)
Constant (Information)	52.20***	54.84***	67.97***	27.52***	59.09***	51.44***	54.99***
	(1.26)	(0.52)	(4.07)	(3.73)	(0.91)	(0.54)	(0.90)
Facility fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
P-value Penalty v Reward	0.447	0.666	0.705	0.545	0.457	0.451	0.483
N respondents	1,363	1,363	1,363	1,363	1,363	1,363	1,363

* p<0.10, ** p<0.05, *** p<0.01. Unadjusted OLS models; s.e. clustered at facility level.

Table A.7: Interaction with awareness that facility participates in NSHIP pilot and understanding of PBF program (percentage points)

	Awareness		Understanding		Understanding if aware and part of PBF	
	(1) Across cases	(2) Across responses	(3) Across cases	(4) Across responses	(5) Across cases	(6) Across responses
Reward	2.96	1.81*	2.94**	2.15**	2.08	2.20
	(1.92)	(1.06)	(1.49)	(0.88)	(4.30)	(2.28)
Penalty	2.75	1.43	3.29**	2.65***	4.41	3.75
	(1.98)	(1.05)	(1.45)	(0.88)	(3.96)	(2.41)
Aware that facility is part of NSHIP	0.11	1.45*				
	(1.58)	(0.87)				
Reward × Aware	1.64	0.96				
	(2.31)	(1.33)				
Penalty × Aware	1.41	1.33				
	(2.34)	(1.32)				
High understanding			0.48	2.17**	1.23	3.39**
			(1.48)	(0.88)	(2.75)	(1.67)
Reward × High understanding			2.35	0.64	2.49	0.33
			(2.13)	(1.29)	(4.66)	(2.53)
Penalty × High understanding			0.95	-0.50	-0.51	-1.95
			(2.13)	(1.29)	(4.35)	(2.64)
Constant (%)	55.42***	56.78***	55.30***	56.86***	55.80***	56.78***
	(1.49)	(0.78)	(1.17)	(0.70)	(2.63)	(1.55)
State FE	Yes	Yes	Yes	Yes	Yes	Yes
NSHIP arms	DFF PFP	DFF PFP	DFF PFP	DFF PFP	PFP	PFP
N respondents	1,182	1,182	1,182	1,182	531	531
R-squared	0.029	0.042	0.031	0.044	0.043	0.058

* p<0.10, ** p<0.05, *** p<0.01. Unadjusted OLS models; s.e. clustered at worker level. High understanding defined as above-median number of correctly named health care services that are incentivized in PBF intervention (median = 6).

Table A.8: Correlation between overall performance on experimental task and screening for dangersigns in direct clinical observation

	Across cases			Across responses		
	(1)	(2)	(3)	(4)	(5)	(6)
Screened for all five danger signs	7.17*			9.57***		
	(3.82)			(3.58)		
Screened for at least one danger sign		7.72**			8.59***	
		(3.20)			(3.00)	
Number of danger signs screened			2.52***			2.97***
			(0.76)			(0.71)
Constant	40.38***	37.26***	36.74***	40.58***	37.45***	36.53***
	(1.77)	(2.49)	(2.21)	(1.66)	(2.34)	(2.06)
Observations	339	339	339	339	339	339

* p<0.10, ** p<0.05, *** p<0.01. OLS models. The five dangers signs to screen for are: fever and severe fatigue; headache and blurred vision; vaginal bleeding; swelling; convulsions. The screening is conducted by the health worker asking the pregnant woman if she has experienced the danger sign in the current pregnancy or in any previous pregnancy.

Table A.9: Percent correct by arm for actions with different incentives

	(1) Correct
Reward	-0.09
	(1.39)
Penalty	1.82
	(1.45)
Payment=50	28.30***
	(1.50)
Payment=100	18.06***
	(1.28)
Payment=200	55.26***
	(1.72)
Payment=300	60.74***
	(2.55)
Reward × Payment=50	8.06***
	(2.23)
Reward × Payment=100	7.23***
	(1.85)
Reward × Payment=200	7.73***
	(2.37)
Reward × Payment=300	6.84**
	(3.40)
Penalty × Payment=50	6.01***
	(2.16)
Penalty × Payment=100	7.01***
	(1.85)
Penalty × Payment=200	3.09
	(2.34)
Penalty × Payment=300	2.84
	(3.43)
Constant (Information × Payment=0)	10.85***
	(0.99)
N actions	21,808
R-squared	0.217

* p<0.10, ** p<0.05, *** p<0.01. OLS models; s.e. clustered at worker level. Listed actions from the three complex cases.

Table A.10: Performance across cases and responses (percentage points)

	Without covariates		With covariates			
	(1) Across cases	(2) Across responses	(3) Across cases	(4) Across responses	(5) Across cases	(6) Across responses
Reward	4.32***	2.04***	0.28***	0.22***	4.46***	1.98***
	(0.97)	(0.45)	(0.06)	(0.05)	(0.96)	(0.45)
Penalty	3.36***	1.88***	0.22***	0.21***	3.41***	1.78***
	(1.00)	(0.45)	(0.07)	(0.05)	(1.00)	(0.44)
Male					0.20	-1.35**
					(1.26)	(0.58)
Age $\geq$ median					-1.82	-0.87
					(1.18)	(0.54)
Doctor or nurse					0.42	2.45***
					(1.83)	(0.79)
Years since qualified $\geq$ median					-0.97	-0.61
					(1.15)	(0.51)
Knowledge $\geq$ median					2.23	1.49**
					(1.58)	(0.64)
Constant (%)	53.21***	56.46***	-0.17***	-0.14***	53.35***	56.60***
	(0.58)	(0.26)	(0.04)	(0.03)	(1.21)	(0.55)
Facility fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
p-values from tests of coefficients						
Penalty v Reward	0.292	0.700	0.292	0.700	0.253	0.637
N respondents	1,363	1,363	1,363	1,363	1,363	1,363
R-squared (overall)	0.018	0.015	0.018	0.015	0.040	0.049

* p<0.10, ** p<0.05, *** p<0.01. OLS models; s.e. clustered at worker level. Omitted category: 15+ years since certification.

Table A.11: Overall performance (percent correct) with covariate interactions

	All cases		By case				
	(1) Across cases	(2) Across responses	(3) Simple 1	(4) Simple 2	(5) Complex 1	(6) Complex 2	(7) Complex 3
A. Interaction with male							
Reward	4.21***	2.58***	3.84	10.09***	1.84**	0.88	4.41***
	(1.13)	(0.53)	(3.75)	(3.45)	(0.83)	(0.57)	(0.96)
Penalty	3.18***	2.50***	1.14	7.50*	2.15**	0.90	4.20***
	(1.19)	(0.51)	(4.02)	(3.83)	(0.93)	(0.60)	(0.93)
Male	-0.41	0.24	-0.42	-2.61	0.30	-2.19*	2.85
	(2.06)	(0.89)	(6.27)	(7.03)	(1.76)	(1.15)	(1.76)
Reward × Male	0.41	-2.32*	6.54	3.55	-2.28	1.23	-6.97***
	(2.64)	(1.24)	(8.00)	(8.46)	(2.41)	(1.44)	(2.25)
Penalty × Male	0.77	-2.94**	6.69	7.42	-3.95	0.80	-7.09***
	(2.96)	(1.27)	(9.09)	(10.06)	(2.60)	(1.69)	(2.28)
Constant (Information)	53.32***	56.39***	68.71***	27.52***	62.24***	52.56***	55.55***
	(0.77)	(0.35)	(2.53)	(2.49)	(0.60)	(0.39)	(0.62)
Facility fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
P-value Penalty v Reward	0.339	0.872	0.463	0.462	0.710	0.967	0.816
N respondents	1,363	1,363	1,363	1,363	1,363	1,363	1,363
B. Interaction with job grade							
Reward	4.74***	2.33***	6.46**	11.18***	1.63**	1.73***	2.71***
	(1.03)	(0.47)	(3.20)	(3.36)	(0.80)	(0.54)	(0.90)
Penalty	3.71***	2.19***	3.27	9.26***	1.55*	1.48**	2.97***
	(1.05)	(0.46)	(3.53)	(3.50)	(0.88)	(0.58)	(0.82)
Doctor or nurse	4.91*	6.42***	-4.89	9.59	7.52**	4.84***	7.47***
	(2.70)	(1.44)	(8.76)	(9.77)	(3.07)	(1.64)	(2.45)
Reward × Doctor or nurse	-5.55	-4.84***	-8.94	-4.52	-5.54*	-6.04***	-2.68
	(3.68)	(1.70)	(13.10)	(12.00)	(3.23)	(1.90)	(3.03)
Penalty × Doctor or nurse	-5.54	-5.54***	-7.90	-3.16	-4.75	-4.84**	-7.03**
	(4.61)	(2.00)	(14.98)	(14.43)	(3.96)	(2.06)	(3.44)
Constant (Information)	52.88***	56.01***	69.04***	26.09***	61.79***	51.64***	55.81***
	(0.64)	(0.28)	(2.02)	(2.08)	(0.50)	(0.33)	(0.52)
Facility fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
P-value Penalty v Reward	0.272	0.768	0.318	0.556	0.917	0.662	0.750
N respondents	1,363	1,363	1,363	1,363	1,363	1,363	1,363

* p<0.10, ** p<0.05, *** p<0.01. OLS models with facility fixed effects and robust standard errors.

Table A.12: Price elasticities of effort

Price	Level in info	%ΔReward-info	%ΔPenalty-info	"E_p" info	$E_p$reward	$E_p$penalty
0	28.30	24.93	19.21	0.14	0.12	0.10
50	18.06	33.34	32.49	0.27	0.50	0.49
100	55.26	13.07	5.45	0.83	0.20	0.08
200	60.74	10.66	4.57	1.52	0.27	0.11

Elasticities from unadjusted OLS models; s.e. clustered at worker level. Based on the three complex cases. While we randomly selected a subset of tasks to be listed or paid, we systematically set higher prices for more salient tasks. Hence, the "price" elasticity in the information arm (where there were no task-specific incentives) reflects the elasticity of effort in response to the salience of a task.

B Instructions and partograph cases

B.1 Instructions

Figure B.1: List provided to participants in "Information" arm

Information arm

Instructions:

We would like you to help us evaluate some partographs.

Here is a list of items that our experts have found important. There might be other things that are not listed here and that are clinically relevant at various stages of labor and delivery.

Note that some items are about NOT doing something because it is UNNECESSARY.

We appreciate your help in examining these partographs and would like to offer 1,750 Naira as a thank-you.

Action	
Refer to secondary facility when necessary	
Measure fetal heart rate at least every 30 minutes	
Monitor contractions every 30 minutes	
Monitor color and consistency of liquor	
Palpate the uterus	
Record all fluids and drugs administered	
Do NOT refer to secondary facility when UNNECESSARY	

Figure B.2: List provided to participants in "Rewards" arm

Reward arm

Instructions:

We would like you to help us evaluate some partographs.

Here is a list of items that our experts have found important. There might be other things that are not listed here and that are clinically relevant at various stages of labor and delivery.

Note that some items are about NOT doing something because it is UNNECESSARY.

We appreciate your help to look at these and would like to offer 1,000 Naira as a thank-you.

As you see, there are numbers next to some items on the list. We will give you those amounts on top of the 1,000 Naira, for every item that you mention and that is clinically indicated in this case. So, if you find some of those items, we will give you more than 1,000 Naira at the end.

These rewards apply to all questions that we'll ask about the partographs.

Action	**Reward** (Naira)
Refer to secondary facility when necessary	300
Measure fetal heart rate at least every 30 minutes	100
Monitor contractions every 30 minutes	50
Monitor color and consistency of liquor	
Palpate the uterus	100
Record all fluids and drugs administered	
Do NOT refer to secondary facility when UNNECESSARY	200

Figure B.3: List provided to participants in "Penalty" arm

Penalty arm

Instructions:

We would like you to help us evaluate some partographs.

Here is a list of items that our experts have found important. There might be other things that are not listed here and that are clinically relevant at various stages of labor and delivery.

Note that some items are about NOT doing something because it is UNNECESSARY.

We appreciate your help to look at these and would like to offer 2,500 Naira as a thank-you.

As you see, there are numbers next to some items on the list. We will subtract those amounts from the 2,500 Naira for every item that you did not mention and that is clinically indicated in this case. So, if you miss some of those items, we will give you less than 2,500 Naira at the end.

These penalties apply to all questions that we'll ask about the partographs.

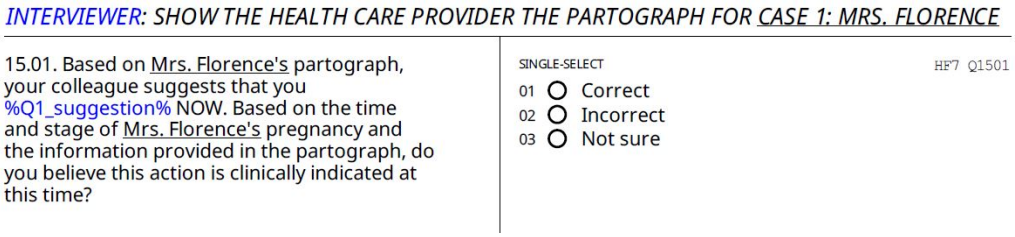
Action	**Penalty** (Naira)
Refer to secondary facility when necessary	300
Measure fetal heart rate at least every 30 minutes	100
Monitor contractions every 30 minutes	50
Monitor color and consistency of liquor	
Palpate the uterus	100
Record all fluids and drugs administered	
Do NOT refer to secondary facility when UNNECESSARY	200

Table B.1: Payout matrix

Action	Reward in Naira	Simple 1	Complex 1	Complex 2	Simple 1	Complex 3	TOTAL
Refer to secondary facility when necessary	300			300			
Measure fetal heart rate at least every 30 minutes	100		100	100		100	
Monitor contractions every 30 minutes	50	50	50	50		50	
Monitor color and consistency of liquor							
Palpate the uterus	100			100			
Record all fluids and drugs administered							
Do NOT refer to secondary facility when UNNECESSARY	200		200		200	200	
Potential max reward / penalty from activities	0	50	350	550	200	350	1,500
Reward arm: Total min payout (only participation incentive)							1,000
Reward arm: Total max payout (participation + activity reward)							2,500
Penalty arm: Total max payout (only participation incentive)							2,500
Penalty arm: Total min payout (participation – activity penalty)							1,000

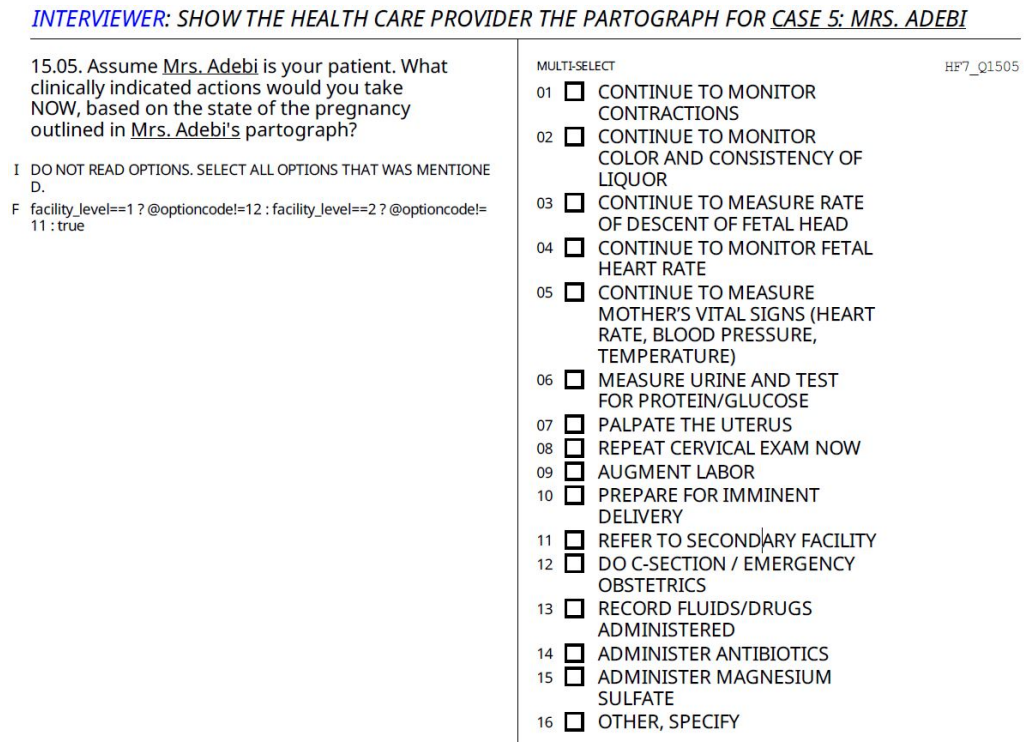
B.2 Partographs

Figure B.4: Example of a simple case: Assessing whether a suggested action is appropriate



Screenshot of CAPI tool for case 1. The action in Q1_suggestion is randomized (within arm) to be (a) "monitor contractions" or (b) "refer to higher level." "Monitor contraction" is an appropriate action, while referring is unnecessary.

Figure B.5: Example of a complex case: Stating appropriate action(s) to be taken



Screenshot of CAPI tool for case 5.

Figure B.6: Partograph case 1

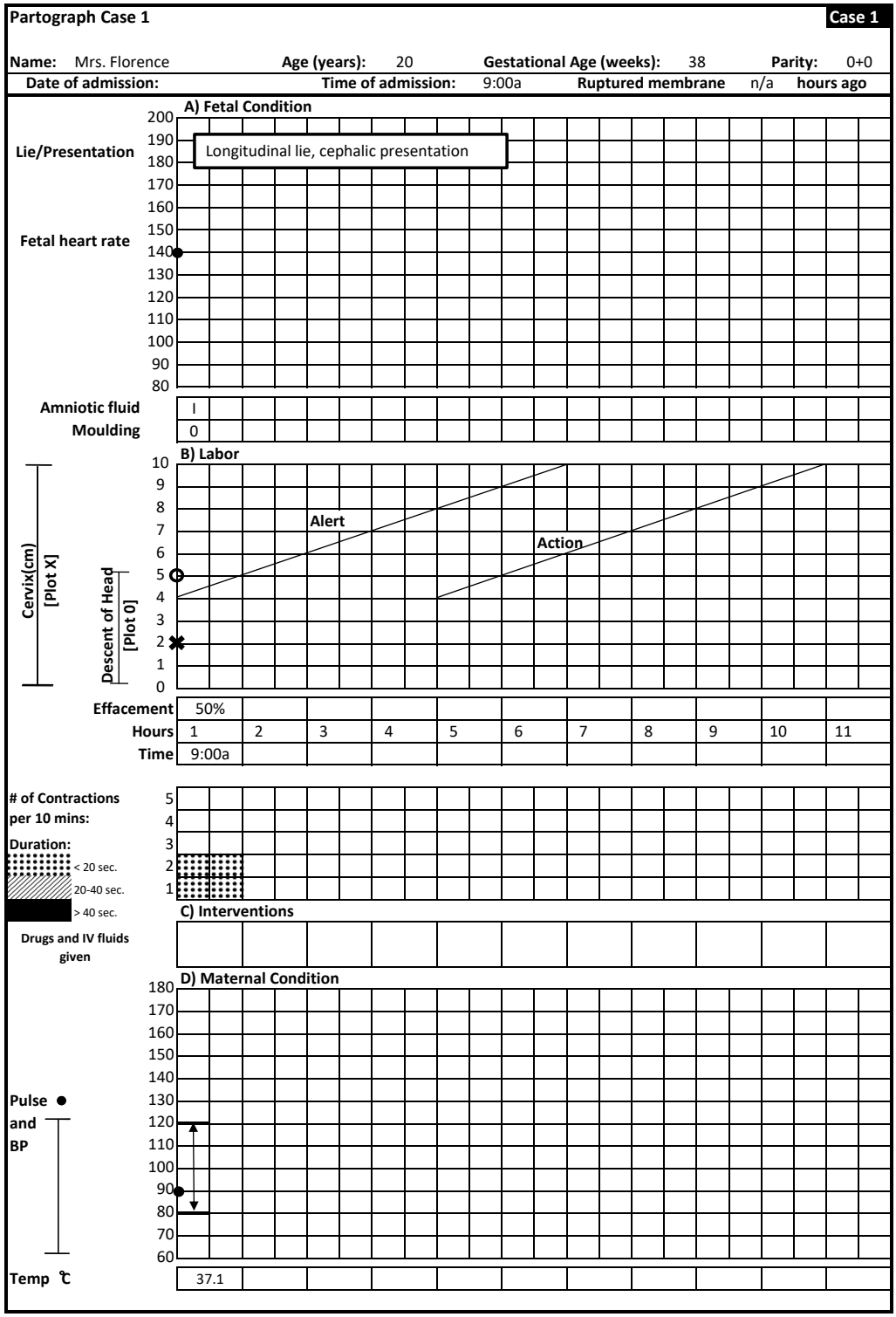


Figure B.7: Partograph case 2

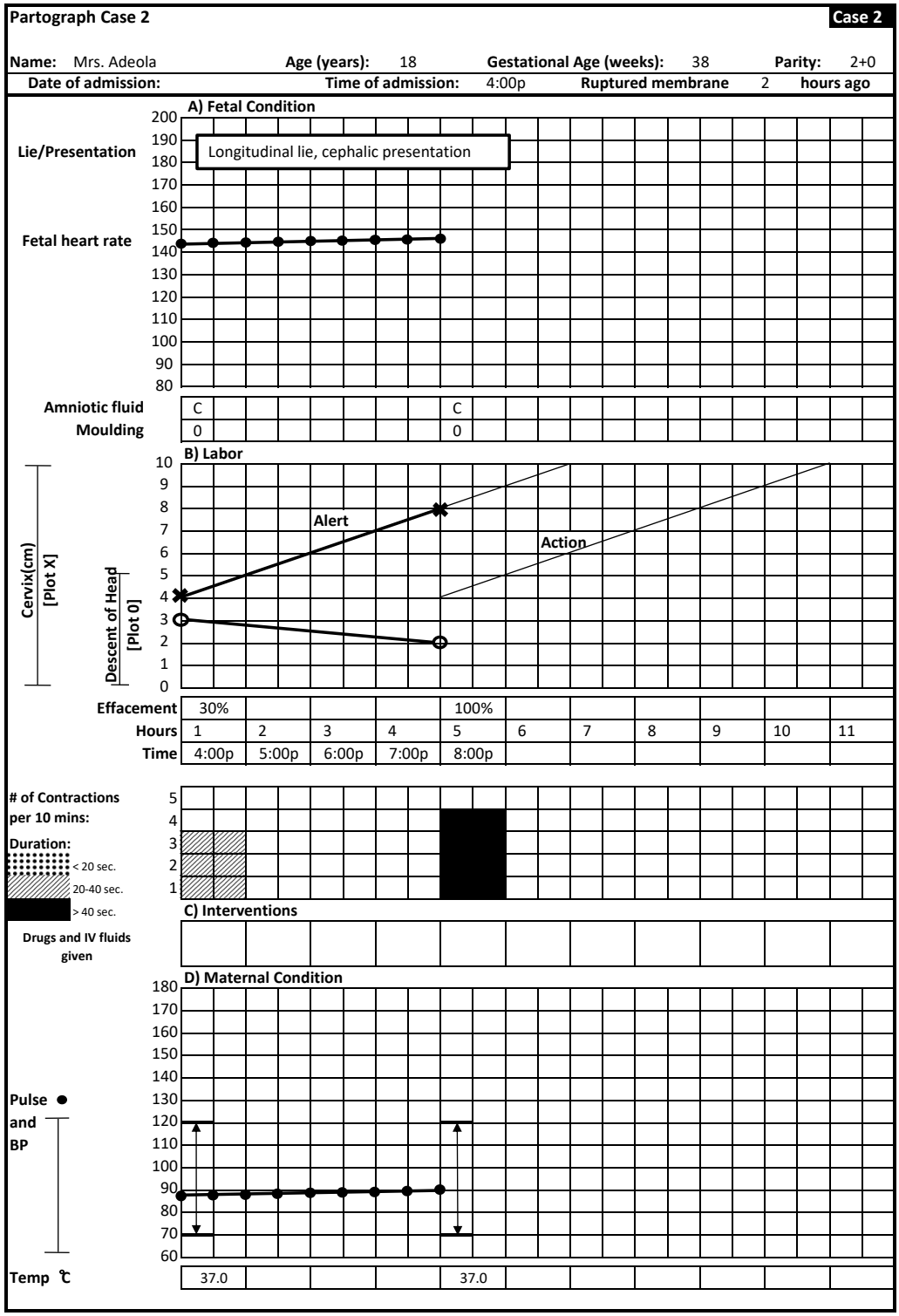


Figure B.8: Partograph case 3

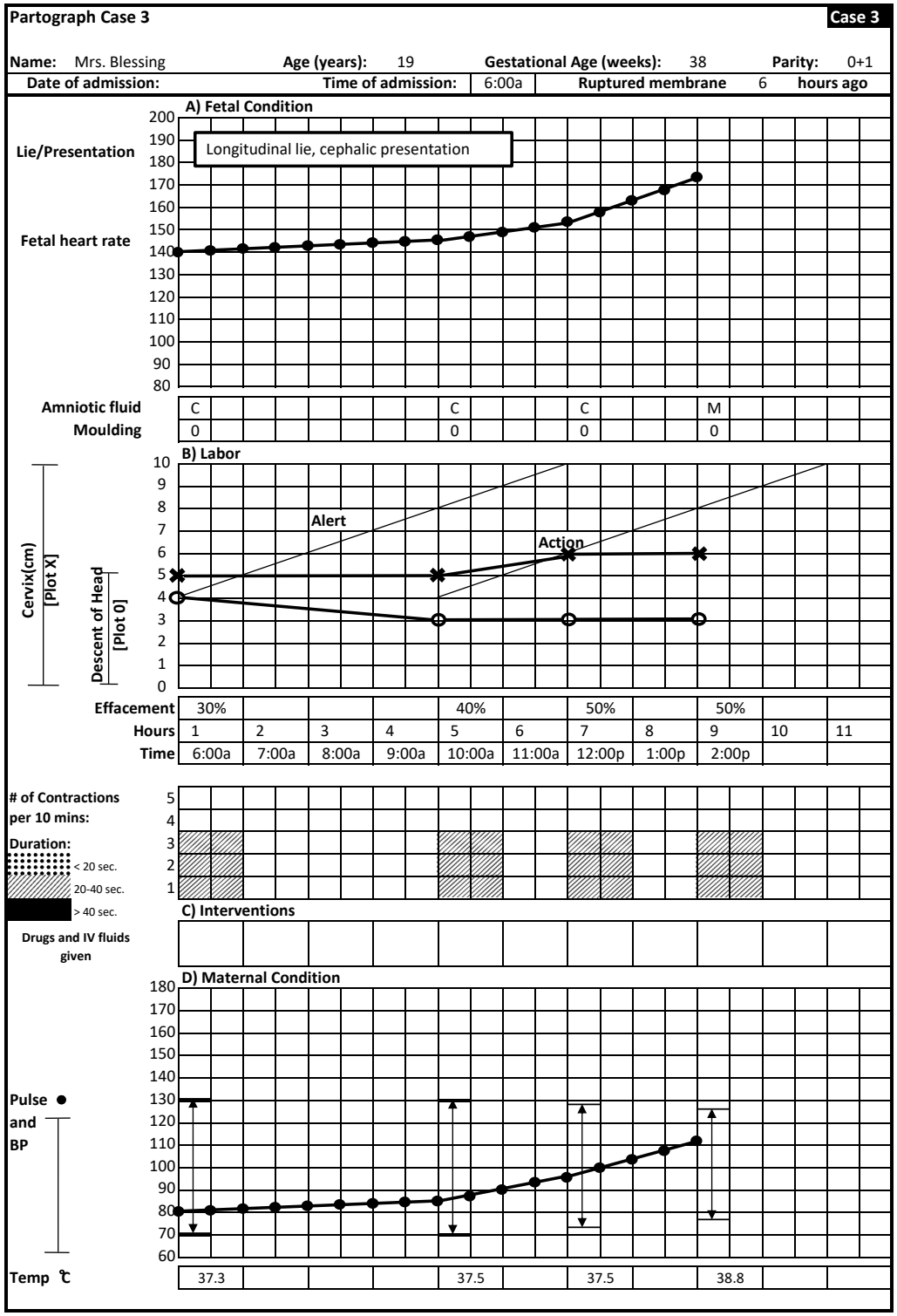


Figure B.9: Partograph case 4

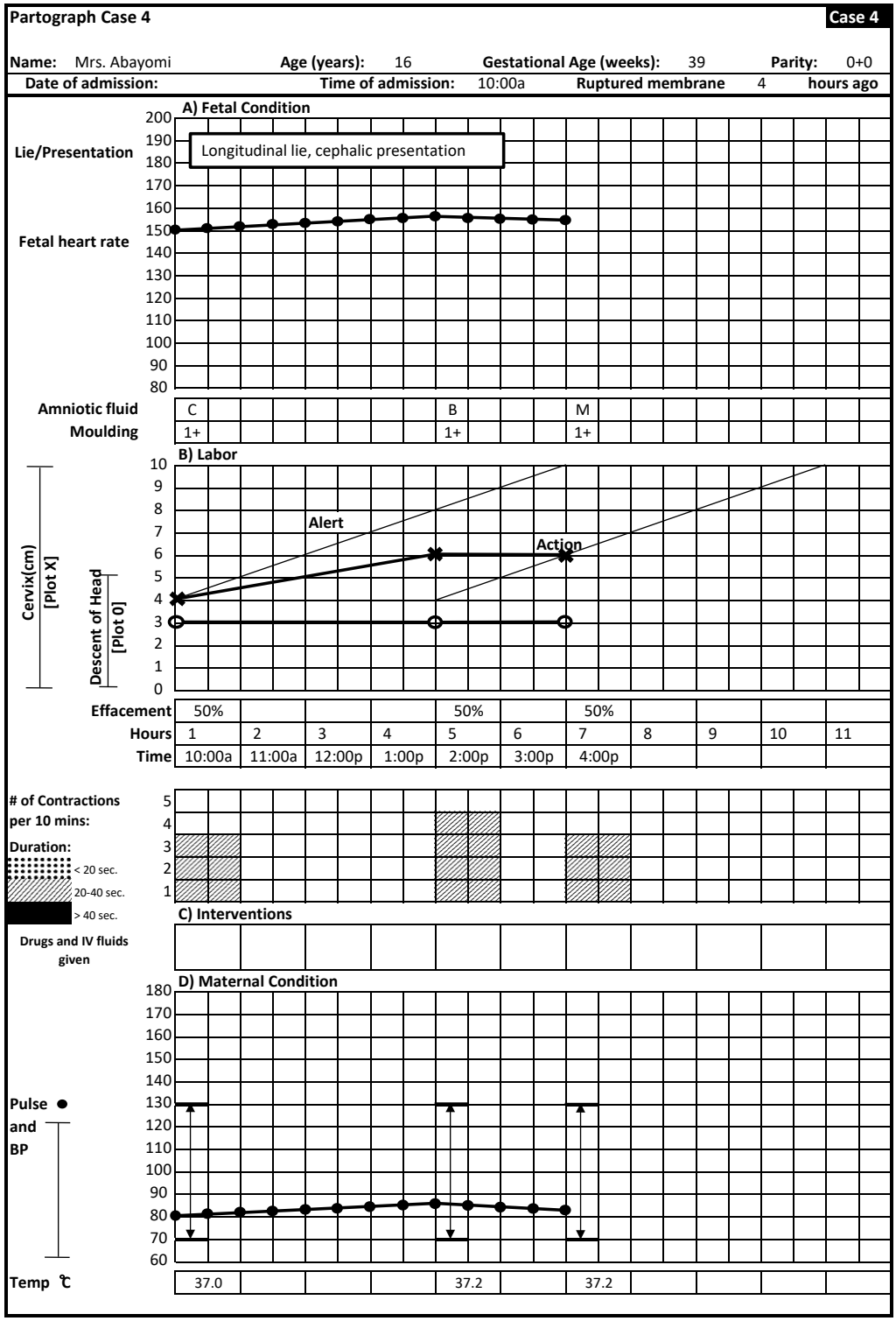
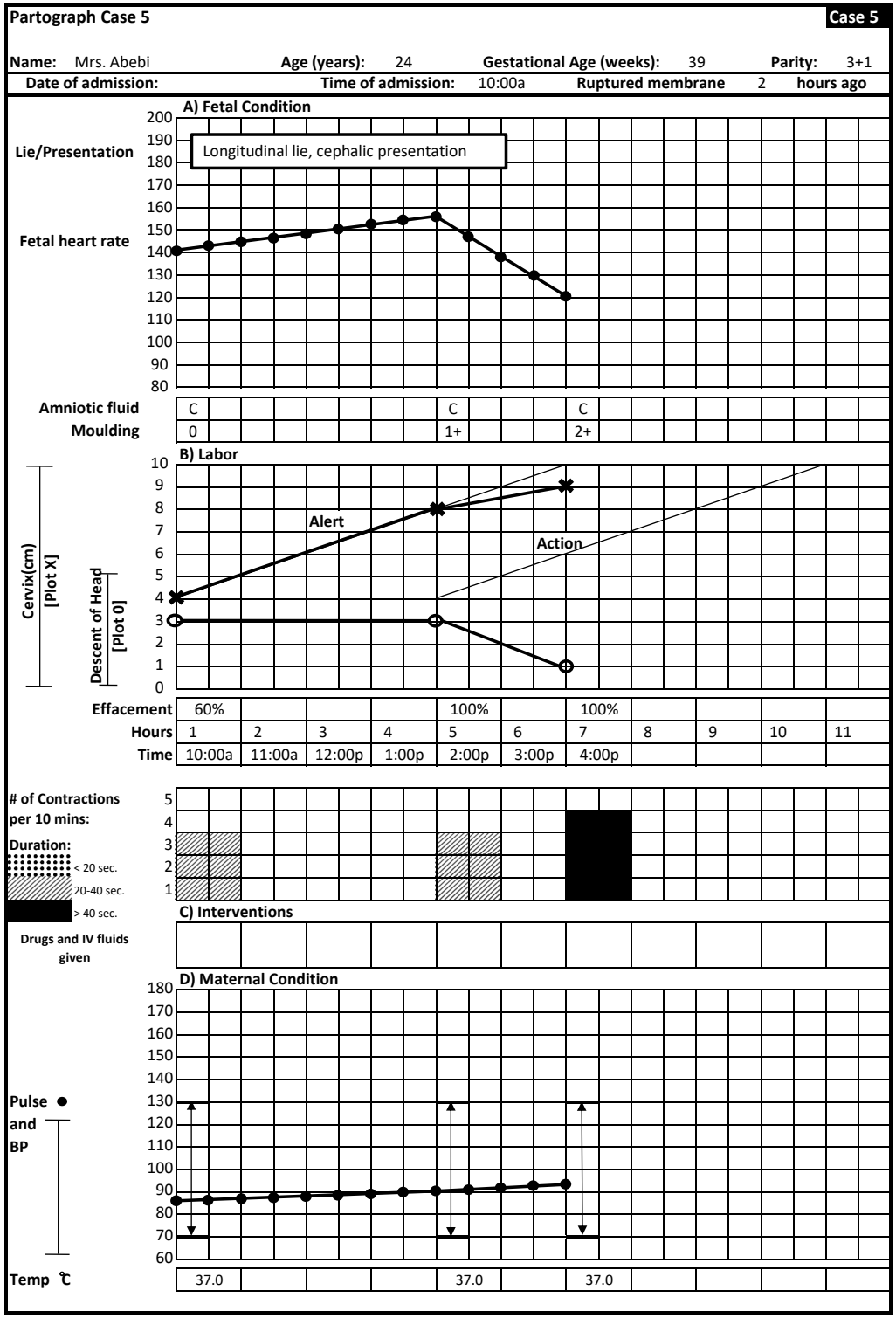


Figure B.10: Partograph case 5



C Deviations from the pre-analysis plan

This annex compares the outcome and covariate variables as described in the pre-analysis plan (PAP) registered in the AEA Trial Registry and as actually implemented.

As summarized in the tables below, the primary analysis does not deviate from the PAP, but the secondary analysis does differ in a few dimensions from what was described in the PAP. First, we had intended to assess the size of the facility catchment area as a key moderator of responses. However, these data were largely missing, thus leaving us unable to complete this portion of the analysis. We had also intended to include the health worker's education level as a covariate but were unable to do so due to a lack of variation in this variable. Initially, we had planned to treat midwives as lower-level professionals, at par with community health workers. But the data show that midwives had training, experience, and salaries commensurate with that of a nurse rather than a community health worker. As a result, we include midwives in the medical professional category, although results are robust to the original specification provided in the PAP. Finally, in lieu of the health worker's education, we explore tenure and experience as additional dimensions along which to assess balance. We added these to the analysis after the PAP was registered. Note that these variables are balanced.

Pre-registered Description	Description as Used in Analysis	Deviation from PAP
Primary Analysis Overview		
AEA description of primary analysis: Proportion, binary measures, and z-scores ofthe share of correct answers. Separately for each of the partographs and combined in indices.	**Actual description of primary analysis**: Included: Proportion and z-scores of correct answers, combined in indices (straight scores). **Design note:** Regression. No covariates. Fixed effects for state and cluster for facility. This is in keeping with the AEA.	**Changes to primary analysis:** • None
Primary Analysis Outcomes		
Outcome 1: Proportion of correct responses. If participants get all responses correct, we will instead use a binary measure of all correct versus not all correct.	**V1: Case score (all responses):** Cases are weighted equally by scoring each case as the proportion of correct responses (given and optional) for that case. Possible range is 0–5.	• None, except that analysis uses two possible calculation methods.
	V2: Response score (all responses): Responses (given and optional) across all cases are weighted equally. Proportion of correct responses is calculated. Possible range is 0–1.	
Outcome 2. Binary measure of "one or more" correct responses.	N/A	• Not used. • Analysis did not use a binary "all correct" measure (or other binary measure) because all-correct scores were very rare or absent, depending on score calculation.
Outcome 3. Z-score of correct responses, where the z-score is calculated as (individual score – mean score) / (standard deviation).	**V1: Standardized case score (all responses):** Same as above, as z-score with mean 0 and sd 1.	• None, except analysis uses two possible calculation methods.
	V2: Standardized response score (all responses): Same as above, as z-score with mean 0 and sd 1.	

Pre-Registered Description	Description as Used in Analysis	Deviation from PAP
Secondary Analysis Outcomes and Covariates		
Outcome: Binary measure whether a specific answer option (a clinical task) was selected.	Same: Yes/no dummies for whether each response is correct.	None
Covariates, of Clinic:		
Level: Binary: primary or secondary	As described in AEA	None
Tercile of the facility's catchment area: Categorical: small, medium, large. Calculated separately for the primary and secondary facilities	Categories were calculated as equal terciles.	• Terciles were not calculated separately for primary and secondary facilities. • **Note:** This was not intentional and can be revisited based on further discussion and/or the specific analysis being performed.
Covariates, of Health Worker:		
Gender: Binary	As described in AEA	None
Age: Continuous	As described in AEA	None
Education Level: Categorical: primary or less, secondary or less, more than secondary	As described in AEA. "More than secondary" included bachelor, master, certificate, diploma, higher national diploma, MBBS (all but "secondary school certificate").	• This covariate is effectively excluded because all but one respondent was categorized as "more than secondary," with the other being "secondary."
Qualification: Categorical: • *medical professional* [doctor, nurse, nurse midwife] • *lower-level medical professional* [midwife, public health nurse, community health officer, community health extension worker, junior community health extension worker] • *other* [e.g., pharmacist, lab technician]	• *medical professional* (doctor or medical officer, nurse, nurse midwife, **midwife**] • *lower-level medical professional* [public health nurse, community health officer, community health extension worker, junior community health extension worker] • *other* [e.g., pharmacist, lab technician]	• Midwife was assigned to "medical professional." • Note: There were no respondents in the "other" category.
Received training in labor/delivery: Categorical: less than 1 year ago, more than 1 year ago, never	As described in AEA	None
N/A (added during analysis, not specified in AEA)	Years since clinical qualification. Continuous and ordinal by <1, 1–2, 2–3, 3–4, 5–9, 10–14, 15–19, 20–29, 30+ years.	• Additional to what is specified in the PAP. Added during analysis.
N/A (added during analysis, not specified in AEA)	Tenure (years at current facility). Continuous and ordinal by <1, 1–2, 2–3, 3–4, 5–9, 10–14, 15–19, 20–29, 30+ years.	• Additional to what is specified in the PAP. Added during analysis.